Hindawi Publishing Corporation
Computational Intelligence and Neuroscience
Volume 2016, Article ID 2381451, 12 pages
http://dx.doi.org/10.1155/2016/2381451

Research Article

Learning to Model Task-Oriented Attention

Xiaochun Zou,[1] Xinbo Zhao,[2] Jian Wang,[2] and Yongjia Yang[2]

[1]*School of Electronics and Information, Northwestern Polytechnical University, Chang'an Campus, P.O. Box 886, Xi'an, Shaanxi 710129, China*
[2]*School of Computer Science, Northwestern Polytechnical University, Chang'an Campus, P.O. Box 886, Xi'an, Shaanxi 710129, China*

Correspondence should be addressed to Xinbo Zhao; xbozhao@nwpu.edu.cn

Received 27 November 2015; Accepted 28 March 2016

Academic Editor: Francesco Camastra

For many applications in graphics, design, and human computer interaction, it is essential to understand where humans look in a scene with a particular task. Models of saliency can be used to predict fixation locations, but a large body of previous saliency models focused on free-viewing task. They are based on bottom-up computation that does not consider task-oriented image semantics and often does not match actual eye movements. To address this problem, we collected eye tracking data of 11 subjects when they performed some particular search task in 1307 images and annotation data of 2,511 segmented objects with fine contours and 8 semantic attributes. Using this database as training and testing examples, we learn a model of saliency based on bottom-up image features and target position feature. Experimental results demonstrate the importance of the target information in the prediction of task-oriented visual attention.

1. Introduction

For many applications in graphics, design, and human computer interaction, it is essential to understand where humans look in a scene with a particular task. For example, an understanding of task-oriented visual attention is useful for automatic object recognition [1], image understanding, or image search [2, 3]. It can be used to direct visual search and foveated image video compression [4, 5] and robot localization [6, 7]. It can also be used in advertising design or implementation of smart cameras [8].

However, it is not easy to simulate task-oriented human visual behavior perfectly by machine. Attention is an abstractive concept, and it needs objective metrics for evaluation. Judging the results of experiments by intuitive observation is not precise because different people might focus on different regions of the same scene, even with task. To solve this issue, eye tracker equipment pieces that can record human eye fixation, saccades, and gazes are routinely used. Investigations of human eye movement data provide more objective ground truth for studies on computational attention models. At the present time, there are over two dozen databases with eye tracking data for both images and videos in the public domain [9], which mainly focus on "free-viewing" eye movements.

Most existing computational visual attention saliency models have often been evaluated against predicting human fixations in free-viewing task, in which some are biologically inspired and based on a bottom-up computational model and others combine both bottom-up image based saliency cues and top-down image semantic dependent cues. Though the models do well qualitatively, the models have limited use because they frequently perform well only in the context-free scenario.

Motivated by this, we make two contributions in this paper. The first is a large database of task-oriented eye tracking experiments with labels and analysis and the second is a supervised learning model of saliency which combines both bottom-up image based saliency cues and task-oriented image semantic dependent cues. Our database consists of eye tracking data from 11 different users across 1307 images. To our knowledge, it is the first time that such an extensive collection of task-oriented eye tracking data is available for quantitative analysis. For a given image, the eye tracking data is used to create a "ground truth" saliency map which represents where viewers actually look with a particular search task. We introduce a set of bottom-up image features and target position features to define salient locations and use a linear support vector machine to train a model of saliency.

We compare the performance of saliency models created with different task-oriented attention and show that our approach performs better in predicting human visual attention regions than MIT model [3], which is one of the best models in predicting context-free human gaze.

The structure of this paper is as follows: Section 2 provides a brief description and discussion of some previous works. Section 3 is devoted to describing the characteristics of the database. In Section 3.1, we present the data collection method, the images, eye tracking data, and ground truth data. Section 3.2 analyzes the properties of our database. The detailed description of our model is in Section 4 that evaluates our approach using the popular saliency model evaluation scores (AUC) with MIT saliency model. The discussion and conclusions are discussed in the last section.

2. Related Work

Attention and saliency play important roles in visual perception. In past few years, more than two dozen of such databases are now available in the public domain. Fixations in Faces (FIFA) [10] were collected from eight subjects performing a 2 s long free-viewing task on 180 color natural images. It demonstrates the fact that faces attract significant visual attention. Subjects were found to fixate on faces with over 80% probability within the first two fixations. The NUSSEF database [11] was compiled from a pool of 758 images and 75 subjects. Each image was presented for 5 seconds and free-viewed by at least 13 subjects. A big feature of this dataset compared with others is that the 758 images in the dataset contain a large number of semantically affective objects/scenes such as expressive faces, nudes, unpleasure concepts, and interactive actions. MIT database from Judd et al. [12] included 1003 images collected from Flickr and LabelMe. Eye movement data were recorded from 15 users who free-view these images for 3 s. In this database, fixations were found around faces, cars, and text. Many fixations are biased towards the center. The DOVES dataset [13] includes 101 natural grayscale images [14]. Eye movements from 29 human observers as they free-view the images were collected. However, all of these databases record "free-viewing" eye movements. In addition, MIT CVCL Search Model Database [15] was recorded to understand task-oriented eye movement patterns of users. Observers were asked to perform a person detection task, and their eye movements were found to be consistent, even when the target was absent from the scene. This database was recorded based on task-oriented attention, but its task is single. So it is necessary to create a content-rich database based on task-oriented attention.

Several visual attention models are directly or indirectly inspired by cognitive concepts which are from psychological or neurophysiological findings. The winner-take-all (WTA) biologically plausible architecture which is related to the Feature Integration Theory is proposed by Koch and Ullman [16]. Built on WTA, Itti et al. [17] first implemented the computational model using a center-surround mechanism and hierarchical structure to predict salient regions. In this model, an image is predecomposed into low-level attributes such as color, intensity, and orientation across several spatial scales. The WTA inference pulls out the position with most conspicuity set of features. Later, Le Meur et al. [18] proposed a bottom-up coherent computational approach based on the structure of the human visual system (HVS), which used contrast sensitivity, perceptual decomposition, visual masking, and center-surround interaction techniques. It extracted features in Krauskopf's color space and implemented saliency in three separate parallel channels: visibility, perceptual grouping, and perception. A feature map is obtained for each channel, and then a unique saliency map is built from the combination of those channels. Based on the isotropic symmetry and radial symmetry operators of Reisfeld et al. [19] and the color symmetry of Heidemann [20], Kootstra et al. [21] developed three symmetry-saliency operators and compared them with human eye tracking data. E. Erdem and A. Erdem [22], Marat et al. [23], and Murray et al. [24] are other models guided by cognitive findings.

Another class of models is derived mathematically. Itti and Baldi [25] defined surprising stimuli as those which significantly change beliefs of an observer. This is modeled in a Bayesian framework by computing the KL divergence between posterior and prior beliefs. Similarly, Zhang et al. [26] proposed SUN (Saliency Using Natural statistics) model in which bottom-up saliency emerges naturally as the self-information of visual features. Bruce and Tsotsos [27] present a model for visual saliency built on a first principles information theoretic formulation dubbed Attention based on Information Maximization (AIM). Avraham and Lindenbaum's work on Esaliency [28] uses a stochastic model to estimate the most probable targets mathematically. Schölkopf et al. [29] proposed the Graph-Based Visual Saliency (GBVS) model, which used a Markovian approach to describe dissimilarity and concentration mass regions. Seo and Milanfar [30] and Liu et al. [31] are two other methods based on mathematical models.

Another class of models computes saliency in the frequency domain. Hou and Zhang [32] proposed Spectral Residual Model (SRM) by relating spectral residual features in spectral domain to the spatial domain. In [28], Avraham and Lindenbaum proposed Esaliency, a stochastic model, to estimate the probability of interest in an image. They roughly segmented the image first and used a graphical model approximation in global considerations to determine which parts are more salient.

Our proposed approaches are related to those models that learn mappings from recorded eye fixations or labeled salient regions. These models use some high-level features obtained from earlier databases and conduct learning mechanisms to determine model parameters. Torralba et al. [33] proposed an attentional guidance approach that combines bottom-up saliency, scene context, and top-down mechanisms to predict image regions likely to be fixated by humans in real-world scenes. Based on a Bayesian framework, the model computes global features by learning the context and structure of images, and the top-down tasks can be implemented in the scene priors. Cerf et al. [34] proposed a model that adds several high-level semantic features such as faces, text, and objects to predict human eye fixations. Judd et al. [12] proposed a learning-based method to predict saliency.

They used 33 features including low-level features such as intensity, color, and orientation; midlevel features such as a horizon line detector; and high-level features such as a face detector and a person detector. The model used a support vector machine (SVM) to train a binary classifier. Zhao and Koch [35] proposed a model similar to that of Itti et al. [17], but with faces as an extra feature. Their model combines feature maps with learned weighting and solves the minimization problem using an active set method. Among the models described above, some focus on adding high-level features to improve predictive performance, while others use machine learning techniques to clarify the relationship between features and their saliency. However, the so-called high-level features are blur concepts and do not encompass all types of environments.

These saliency models have been used to characterize RoIs in free-viewing task, but their use in particular task has remained very limited. Recent results suggest that, during task-oriented visual attention, in which subjects are asked to find a particular target in a display, top-down processes play a dominant role in the guidance of eye movements [36–40]. However, the so-called top-down features are blur concepts and do not encompass all types of environments. Here, we exploit more informative concepts including low-level, target location, and center bias, using machine learning for eye fixation prediction.

3. Database of Eye Tracking Data

We collected a large database of eye tracking data to allow large-scale quantitative analysis of fixation points and gaze paths and to provide ground truth data for saliency model research [41]. Compared with several eye tracking datasets that are publicly available, the main motivation of our new dataset is for studying task-oriented visual attention, that is, where observers look while deciding whether a scene contains a target.

3.1. Data Gathering Protocol

3.1.1. Participants. Fifteen participants, undergraduate and graduate volunteers aged 19–32 years ($\mu = 23.3$, $\sigma = 38.4$) with uncorrected and corrected normal eyesight, voluntarily joined this experiment. All the participants were from the Northwestern Polytechnical University.

3.1.2. Apparatus. Tobii TX300 eye tracker was used to record eye movements. We set the sampling frequency to 300 Hz. The eye tracker tolerates a certain extent of head movements, which allows the subjects to move freely and naturally in front of the stimulus. Freedom of head movement is at 65 cm, 37 × 17 (width × height), where at least one eye is within the eye tracker's field of view. Max head movement speed 50 cm/s stimuli were presented on a 23-inch wide screen TFT monitor. The screen size was 50.5 cm × 28.5 cm. Its screen response time was typically 5 ms and its resolution was set to 1920 × 1080.

3.1.3. Materials. We randomly selected 1307 images from VOC2012 as the stimuli. The longest dimension (could be either width or height) of each image was 500 pixels and the other dimension ranged from 213 to 500 pixels. The images contained eight categories, namely, airplane, motorbike, bottle, car, chair, dog, horse, and person.

3.1.4. Procedure. The 1307 images were separated into eight groups. Each group contained 100 images from the same categories and 70 images from the other categories (10 images were selected from each of these categories). All subjects sat at a distance of approximately 65 cm from the screen in a relatively quiet room. The images from each group were presented randomly with their original size in the middle of screen. Before the test, a five-point target display was used for calibration. To ensure high-quality tracking results, we checked the calibration accuracy after each of the groups. If the accuracy of the eye tracker was within about 1° visual angle, the subjects can continue the next group. Otherwise, the calibration will be carried out again. Subjects will be given different instructions for each of the groups. For example, for airplane group, subjects would be asked to find airplane in each picture, while a picture may have zero, one, or more airplanes. Subjects should find airplanes as more as possible in one image and switch to the next one through hitting the space key. To encourage the subjects to concentrate on looking for the target, we took two measures to improve authenticity of test. On the one hand, each group (above-mentioned eight groups) was equally divided into three small subsets. Subjects will spend less time to view the small subsets and pay more attention to the stimuli. On the other hand, after each subset, the subjects took a 2 min break and did a memory test: how many airplanes did you find?

3.2. Analysis of Dataset

3.2.1. Consistency. In our dataset, for the target-present images, all subjects fixate on the same locations, while, in target-absent image, subjects' fixations are dispersed all over the image. We analyze this consistency of human fixations over an image by measuring the entropy of the average continuous saliency map across subjects. Though the original images were of varying aspect ratios, we resized them to 200 × 200 pixel images before calculating entropy. Figure 1(c) shows a histogram of the entropies of the images in our database. It also shows a sample of 12 saliency maps (shown in Figures 1(a) and 1(b)) with lowest and highest entropy and their corresponding images.

3.2.2. Center Bias. Our data indicates a strong bias for human fixations to be near the center of the image, as is consistent with previously analyzed eye tracking datasets [12, 42]. Figure 2 shows the average human saliency map separately from the dog and chair category, which have the strongest and weakest center bias. In the dog category, 57% of the gaze points lie within the center 11% of the image, and 80% of the gaze points lie within the center 25% of the image. In the chair category, 29% of the gaze points lie within the

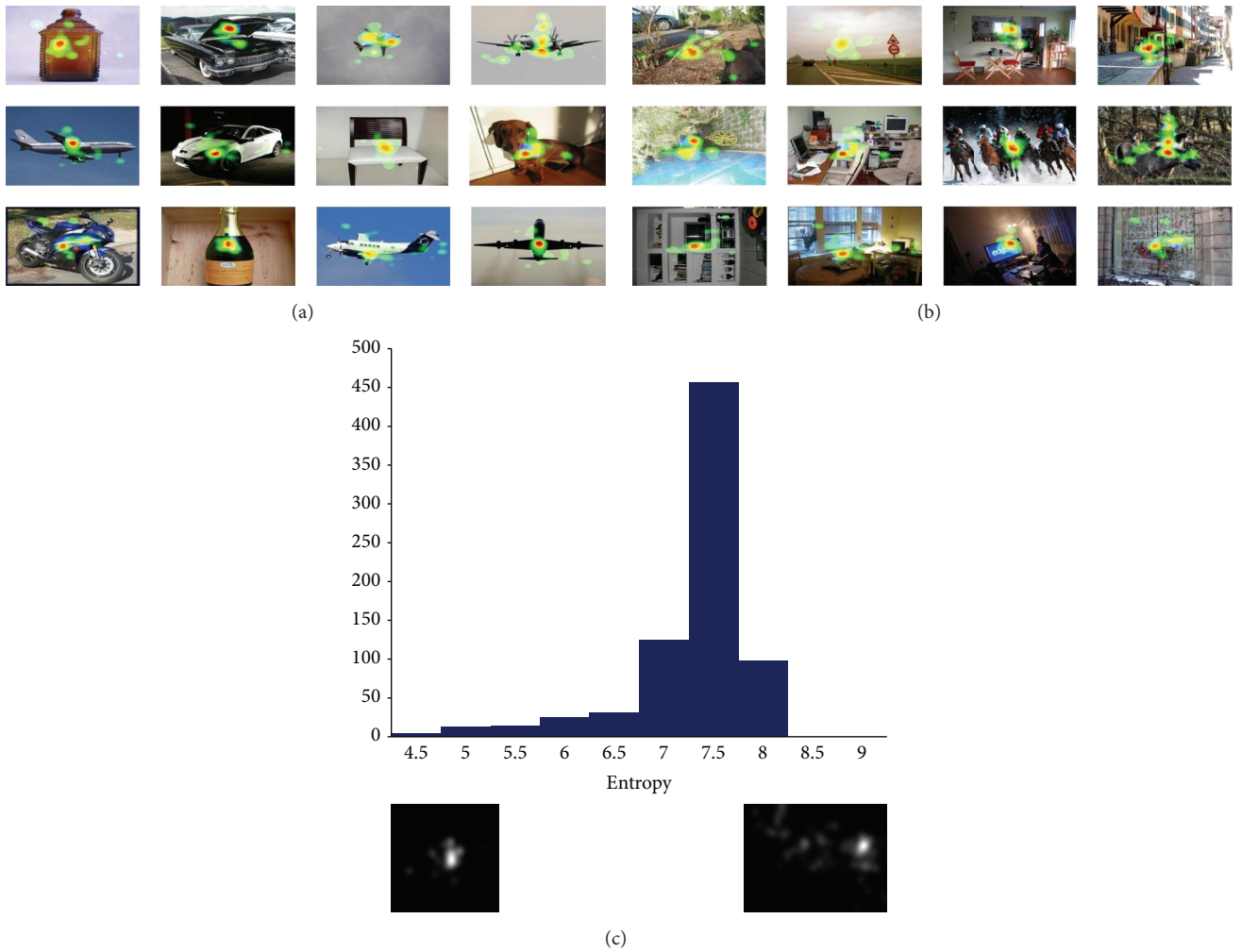


Figure 1: ((a) and (b)) The heat map made from subjects gaze points with low and high entropy. If the image has high entropy, it usually contains more objects. (c) A histogram of the saliency map entropies.

center 11% of the image, and 49% of the gaze points lie within the center 25% of the image.

There are several hypotheses for the root cause of center bias. In our test, the main reason is that people tend to place object or interesting things near the center of an image when taking a picture (the so-called photographer bias). To test this notion, we separately analyze percent of target gaze points, which are gaze points located on the target object within the center 11% and 25% of the dog and chair category. Obviously, in the dog category percentage of target gaze points in center area is more than that in the chair category. This difference has been attributed to the fact that target object mainly located on the center of images in dog category but was distributed in the whole image in chair category.

3.2.3. Agreement among Observers. In this paragraph, we evaluate agreement of the fixation positions among observers. Analysis of the eye movement patterns across observers showed that the fixations were strongly constrained by the search task and the scene context. To evaluate quantitatively the agreement among observers, we studied the human interobserver (IO) model to predict eye fixations, under the same experimental conditions. The IO model outputs, for a given stimulus, a map built by integrating eye fixations from subjects other than the one under test while they watched that stimulus. Then the map was used to predict fixations of the excluded subject. Finally, we use the evaluation of IO model performance to evaluate the agreement among observers.

Using the area under the ROC curve (AUC) as the score, the IO model's map is treated as a binary classifier on every pixel in the image. Pixels with larger values than a threshold are classified as fixated while the rest of pixels are classified as nonfixated. Human fixations are used as ground truth. By varying the threshold, the ROC curve is drawn as the false positive rate versus true positive rate, and the area under this curve indicates how well the saliency map predicts actual human eye fixations.

We separately computed the IO model over 8 categories from our dataset and select the mean value as the result. Table 1 shows the mean value of AUC scores of models.

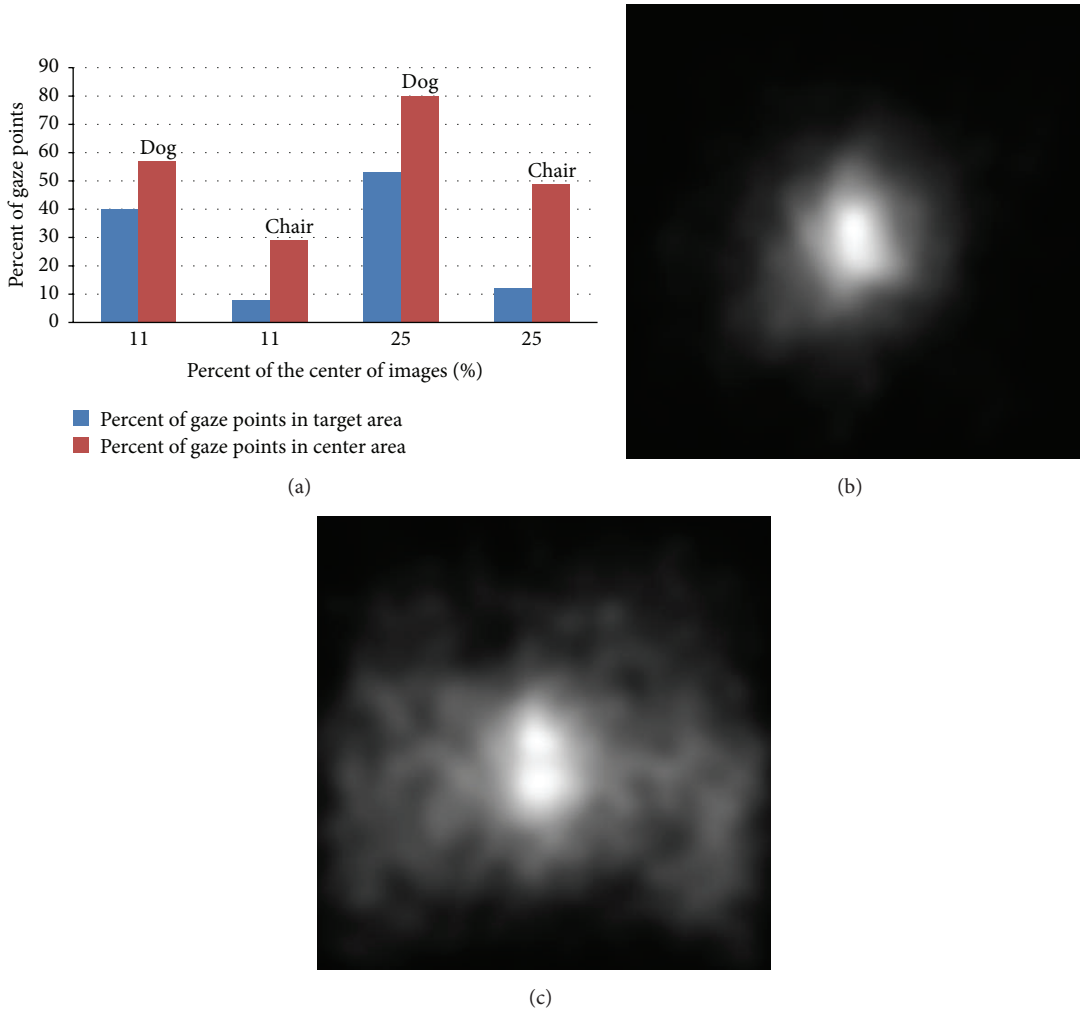


Figure 2: (a) The percentage of gaze points within the center 11% and 25% of the images, which is displayed by blue. Meanwhile, red shows the percentage of target gaze points. Obviously, in the dog category, the percentage of target gaze points is more than chair category. ((b) and (c)) Dog's and chair's average saliency map containing all the gaze points, which indicates a bias to the center of the image.

Table 1: Intersubject agreement for target-present and target-absent.

Group name	Target-present	Target-absent
Airplane	0.90	0.90
Bottle	0.87	0.87
Car	0.86	0.86
Chair	0.83	0.84
Dog	0.95	0.95
Horse	0.94	0.94
Motorbike	0.93	0.93
Person	0.92	0.93
Average	0.90	0.92

The results show that observers are very consistent with one another on the fixated locations in the target-present and target-absent conditions (over 85% in each case). On average, the agreement among observers is higher when the target is present than absent. This suggests that locations fixed by observers in target-present image are driven by the target location.

3.2.4. Gaze Points in Each Stimulus. The task of counting target objects within picture is similar to an exhaustive visual search task. In our design, each scene could contain up to 4 targets. Target size was not prespecified and varied among the stimuli set. Under these circumstances, we expected observers to exhaustively search each scene, regardless of the true number of targets present. Figure 3 shows the average number of the total of gaze points of each stimulus in every group. Unexpectedly, the count of fixations in the target-present is obviously more than target-absent.

To analyze the fixation position in the target-present images, we compare the percentage of human fixation that falls within the target object and the center area. In the first case, we apply the ground truth segmentation as the target object's area. In the second case, we calculate the percentage of human fixations located within the center 2%, 11%, 25%,

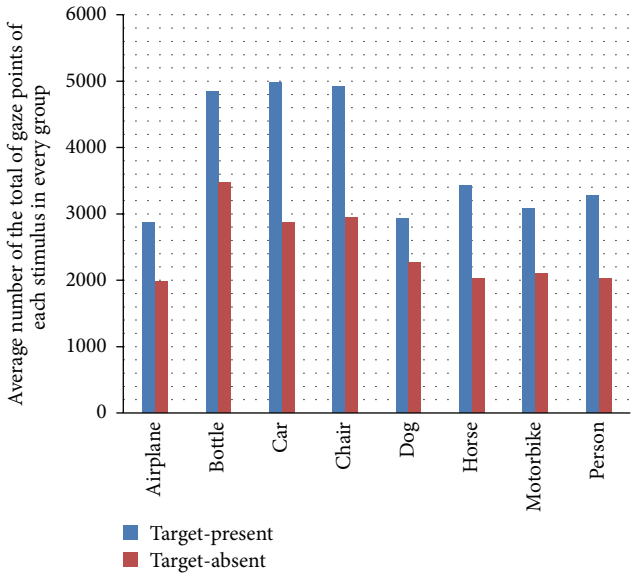


Figure 3: Average number of the total of gaze points of each stimulus in every group.

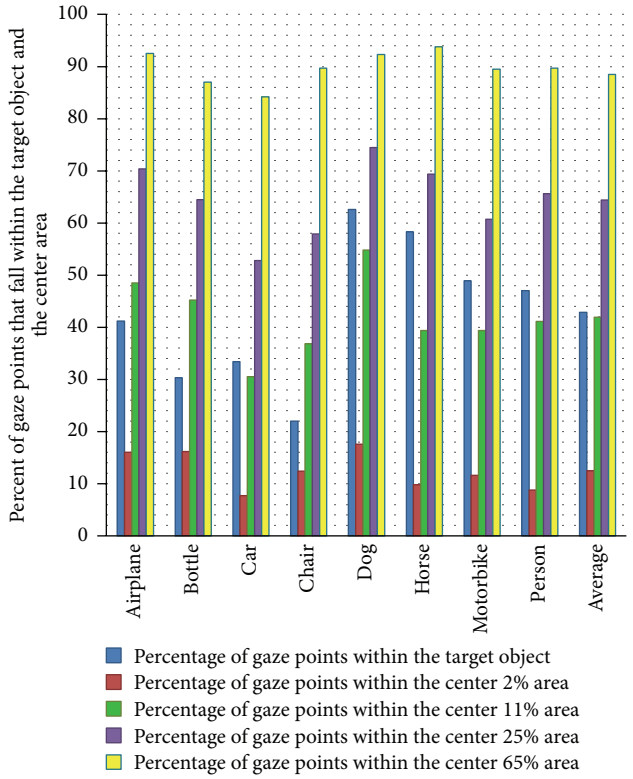


Figure 4: Percent of gaze points that fall within the target object and the center area.

and 65% of the image. Figure 4 summarizes the results. First of all in two cases, the percentages both are above chance level. The differences seen in Figure 4 are statistically significant: the center 25% of the image better attracts human fixations than the target object area. This effect was mostly driven by subject's sidelong glance, for which human fixations are always around target object. But even so, the graphs in Figure 4 clearly indicate that the location of target object (the center area) and the area of target object will attract human fixations.

3.2.5. Objects of Interest. According to Judd et al. [12, 42], if stimuli have one or more humans, gaze points should mainly locate on the human faces. However, in our test, this situation is not similar.

Figure 5 shows heat map of stimuli in which have one or more humans. From Figure 5, we can know the following:

(a) For one stimulus, it has different heat map in different situation.

(b) If the human is the target object, many gaze points still locate on the human face.

(c) When subjects search target in the stimuli, they can ignore the other objects and pay all attention to the target object.

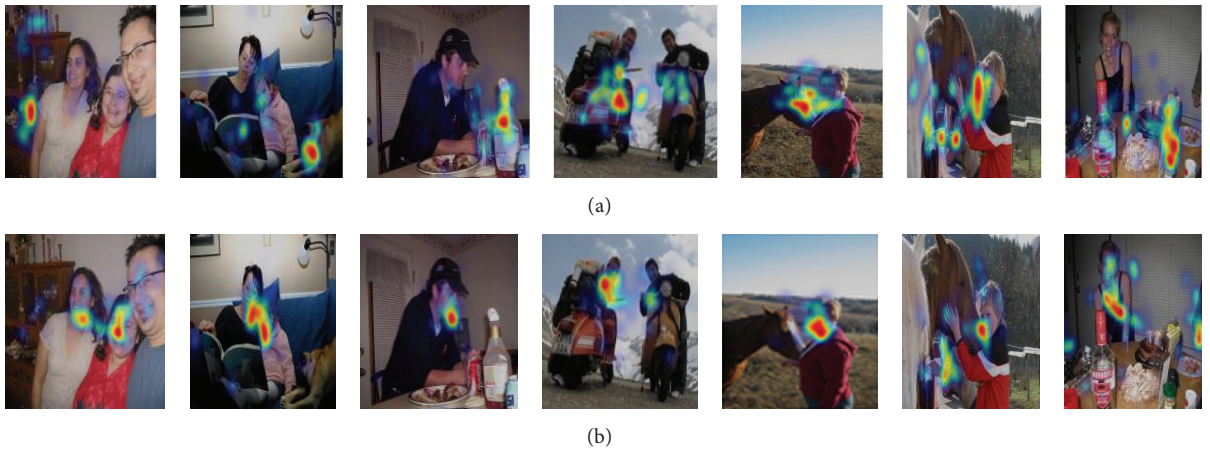
From what we have discussed above, we know that in our test whether some object is of interest depends on the task.

4. Learning-Based Saliency Model

In contrast to previous computational models that combine a lot of biologically plausible filters together to estimate visual saliency, we use a learning approach to train a classifier directly from human eye tracking data. For each image, we precomputed the feature maps for every pixel of the image resized to 200 × 200 and used the maps to train our model. Figure 6 shows the feature maps. Through analyzing our dataset, we promoted low-level, high-level, and center prior features.

Low-level features, intensity, orientation, and color contrast have long been seen as significant features for bottom-up saliency. We include the three channels corresponding to these image features as calculated by Itti and Koch's saliency method [43]. Regarding high-level features, according to our data analysis, we found that humans gaze points always located on target object. So we used the location of target object as the high-level features. Firstly, bounding boxes around objects were labeled and we used them as the target object's area. Secondly, in the boxes, we used the distance of every pixel to the center of box instead of the pixel. Finally, out of boxes, we used zero instead of the pixel. Center bias, when humans take pictures, they naturally frame an object of interest near the center of the image. For this reason, we include a feature which indicates the distance to center of each pixel [12].

To evaluate our model, we followed the 5-fold cross validation method. The method partitions the database into five subsets randomly, each with M images. Every subset is selected sequentially as a test set and the remainders serve as the training set. Each time we trained the model from 4 parts and tested it over the remaining part. Results are then averaged over all partitions. From the ground truth gaze point



(a)

(b)

Figure 5: The figure shows the heat map of stimuli. (a) It shows the target-present's heat map but human is not target object. (b) It shows the target-present's heat map but human is target object.

Table 2: The table shows the average (Avg) and the corresponding standard deviations (STD) of the weight of attribute in each category. For every category, the bold weight is the first and the second is italic weight.

Category	Color		Intensity		Orientation		Target		Center bias	
	Avg	STD	Avg	STD	Avg	STD	Avg	STD	Avg	STD
Airplane	0.0319	0.00005	−0.0154	0.00002	0.0098	0.00002	*0.1201*	0.00012	**−0.4344**	0.00025
Bottle	0.0346	0.00006	0.0424	0.00006	0.0294	0.00004	*0.1206*	0.00009	**−0.3586**	0.00019
Car	0.0073	0.00001	0.0112	0.00002	−0.0159	0.00002	**0.2575**	0.00016	*−0.2418*	0.00012
Chair	0.0234	0.00003	0.0578	0.00006	0.1002	0.00011	**0.2766**	0.00013	*−0.1348*	0.00008
Dog	0.0066	0.00001	0.0075	0.00001	0.0848	0.00006	*0.1065*	0.00008	**−0.4556**	0.00024
Horse	0.0241	0.00004	−0.0004	0.00000	0.0240	0.00003	*0.1445*	0.00011	**−0.3182**	0.00031
Motorbike	−0.0088	0.00002	0.0166	0.00002	0.0276	0.00004	*0.2001*	0.00015	**−0.2733**	0.00025
Person	−0.0131	0.00003	−0.0291	0.00003	0.0638	0.00006	*0.1241*	0.00007	**−0.3159**	0.00027

map of each image, 20 pixels were randomly sampled from the top 20% salient locations, and 20 pixels were sampled from the bottom 70% salient locations to yield a training set of 3200 positive samples and 3200 negative samples. The purpose of choosing a 1 : 1 sampling ratio is to balance the distributions of positive and negative sample pixels in the same image. We chose samples from top 20% and bottom 70% in order to have samples that were strongly positive and strongly negative. The training samples were normalized to have zero mean and unit variance. The same parameters were used to normalize the test set.

We used the linear support vector machine [44] to train the model which was first used to learn the weight of each low-level, high-level, and center prior attribute in determining the significance in attention allocation. We used models with linear kernels because they are faster to compute, and the resulting weights of attributes are intuitive to understand. For each group, the average (Avg) and the corresponding standard deviations (STD) across the number of experiment executions of the learned weight of each attribute are shown in Table 2. It is clear that the attribute of center bias and the location of target object have the higher weight than others. Obviously, in the dog group, the weight of center bias is stronger than others. However, in the chair group, the weight of the location target object is stronger than others. For this phenomenon, the flowing may be critical. The areas of target object may contribute to the phenomenon. But we do not know the detailed relations. The weight of attribute also agrees with previous finding in figure-ground perception that, during visual search tasks, in which subjects are asked to find a particular target in a display, top-down processes play a dominant role in the guidance of eye movements.

5. Evaluation

To measure performance of saliency models, we performed comparisons of our models with the MIT model [3] which is one of the best models in predicting context-free human gaze. The model incorporated bottom-up saliency and high-level image semantics and works well in predicting saliency in a free-viewing context. To make the result comparable, the MIT model is trained on the same training set as our method. Figure 7 shows heat maps of our model and the compared model. This is result for one image in each group. We conducted our experiment on 160 images randomly selected.

Figure 8 shows Receiver Operating Characteristic (ROC) for our model and MIT model. These curves show the

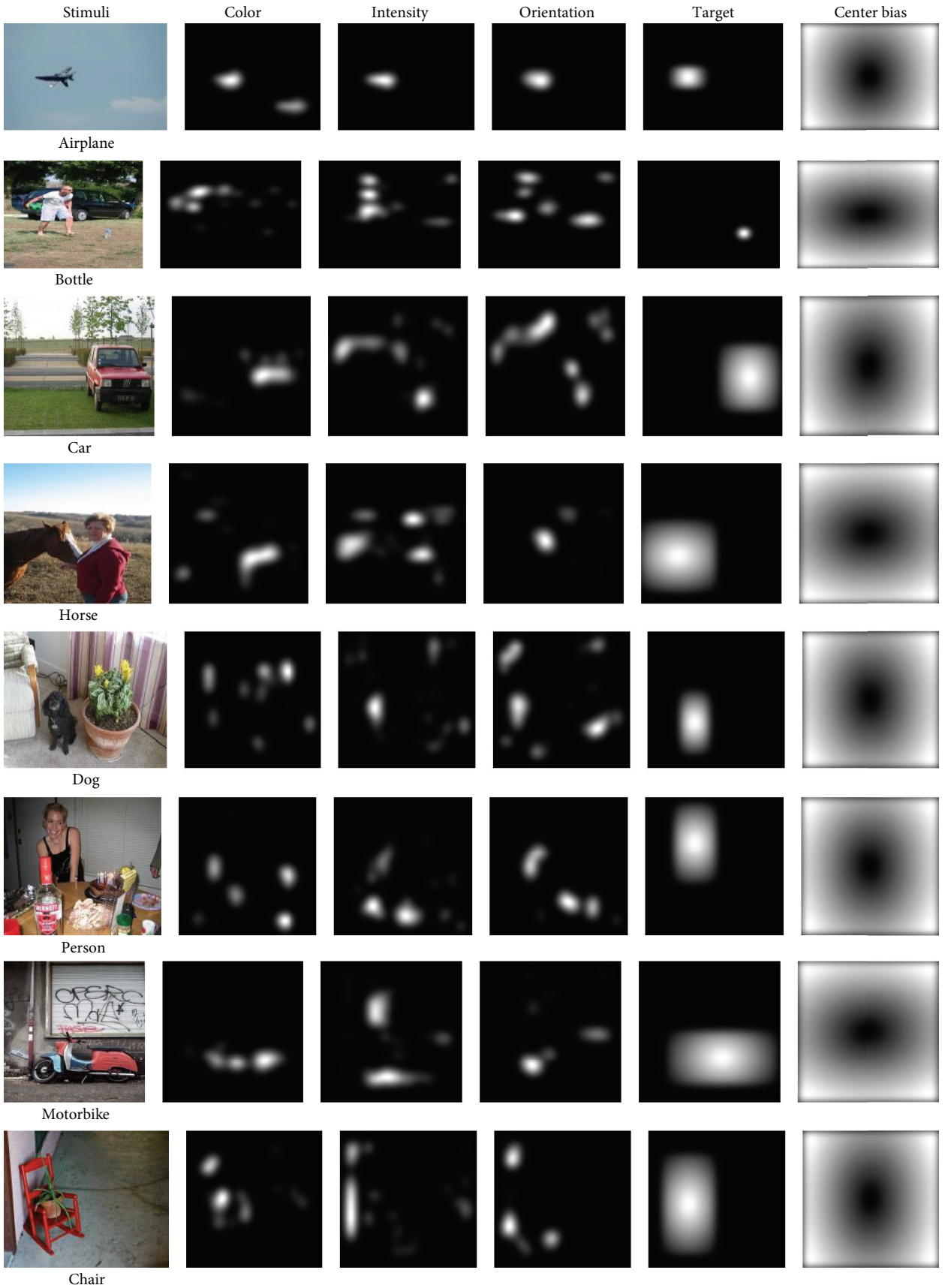


Figure 6: The figure shows the low-level feature maps such as color, intensity, orientation, and high-level feature maps such as the location of target object, finally, center-bias feature map.

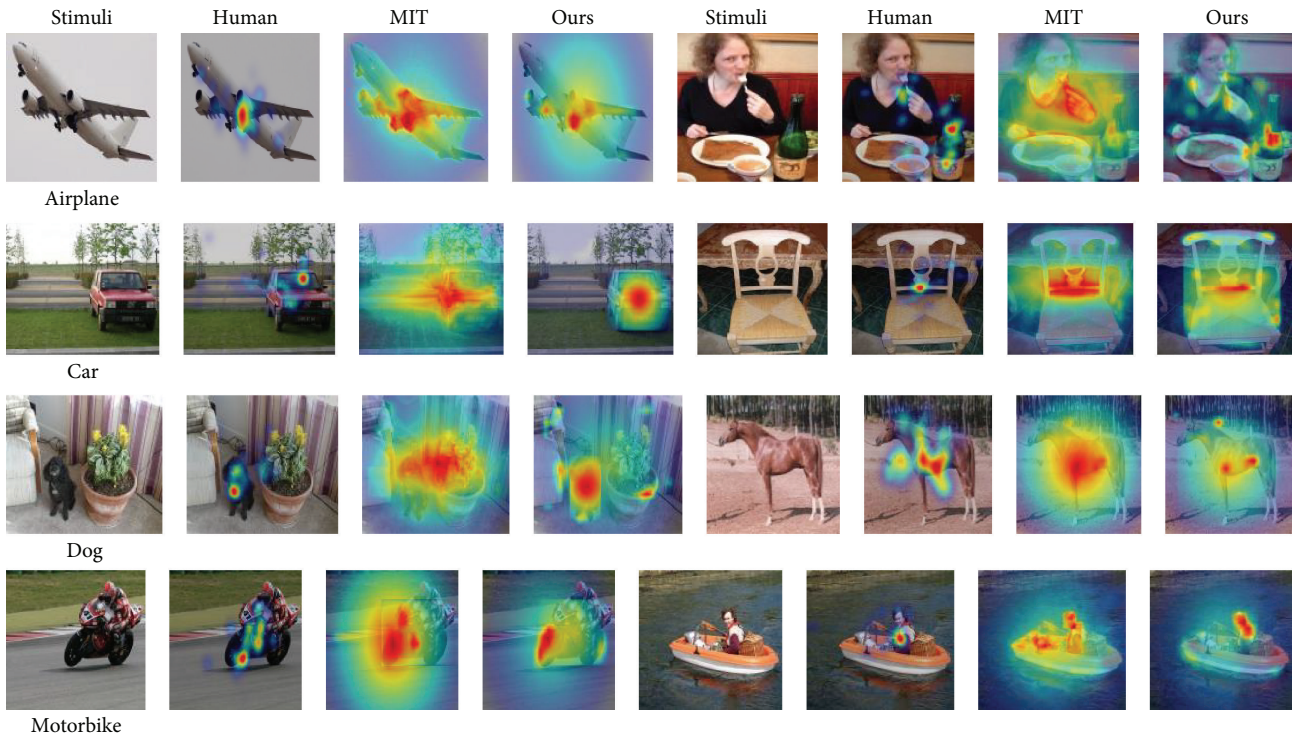


Figure 7: The figure shows the heat maps, which are generated by our model and MIT model. They were trained by the same gaze points and used the same training method.

Table 3: The table shows the average (Avg) and the corresponding standard deviations (STD) of the AUC in each category.

Model	Category							
	Airplane	Bottle	Car	Chair	Dog	Horse	Motorbike	Person
MIT								
Avg	0.8572	0.7881	0.7865	0.8152	0.8639	0.8583	0.8563	0.7962
STD	0.0016	0.0012	0.001	0.0015	0.0006	0.0008	0.0013	0.0011
Ours								
Avg	0.8635	0.8566	0.8873	0.9015	0.8665	0.8663	0.8563	0.8893
STD	0.0012	0.0006	0.0005	0.0004	0.0006	0.0009	0.0007	0.0007

proportion of gaze points that fall within the saliency map predicated by saliency model (detection rate) in relation to the proportion of the image area selected by the saliency map (false alarm rate). Our saliency models were generated by a weighted linear combination of the feature maps using the learned weights of each attribute. It shows how well the gaze points of each subject can be predicted by saliency model. For each category, we calculate the average (Avg) and the corresponding standard deviations (STD) across the number of experiment executions of the area under the ROC curve (AUC), which is shown in Table 3.

It can be seen that, for the MIT model, the performance is not always well; however, our model is better than MIT. For example, in bottle, car, and chair category, MIT model predicted gaze points regions with lower accuracy (AUC = 0.7881, AUC = 0.7865, and AUC = 0.8152) than our models (AUC = 0.8566, AUC = 0.8873, and AUC = 0.9015). From Table 3, we know that the weight of location of target object is first in the car and chair category. So, the promotion of accuracy mainly results from target guidance factor. However, even our model could not compete with human agreement.

6. Discussions and Conclusions

According to Figure 8 and Table 3, it is obviously shown that, for the bottle, car, and chair category, MIT model has lower performance, while our model has larger better performance than it. The main factor is that in these categories target object is small or not salient, so when subjects are free-viewing, they are not saliency map. However, in the task-oriented attention, they become the saliency map; that is why free-viewing model is not appropriately task-oriented.

As we all know several recent datasets [10–12, 45] all set the free-viewing time to 2–5 s per image. In our paradigm, the time was given to the subjects, which is mostly motivated by the following factors. If the viewing duration is too short,

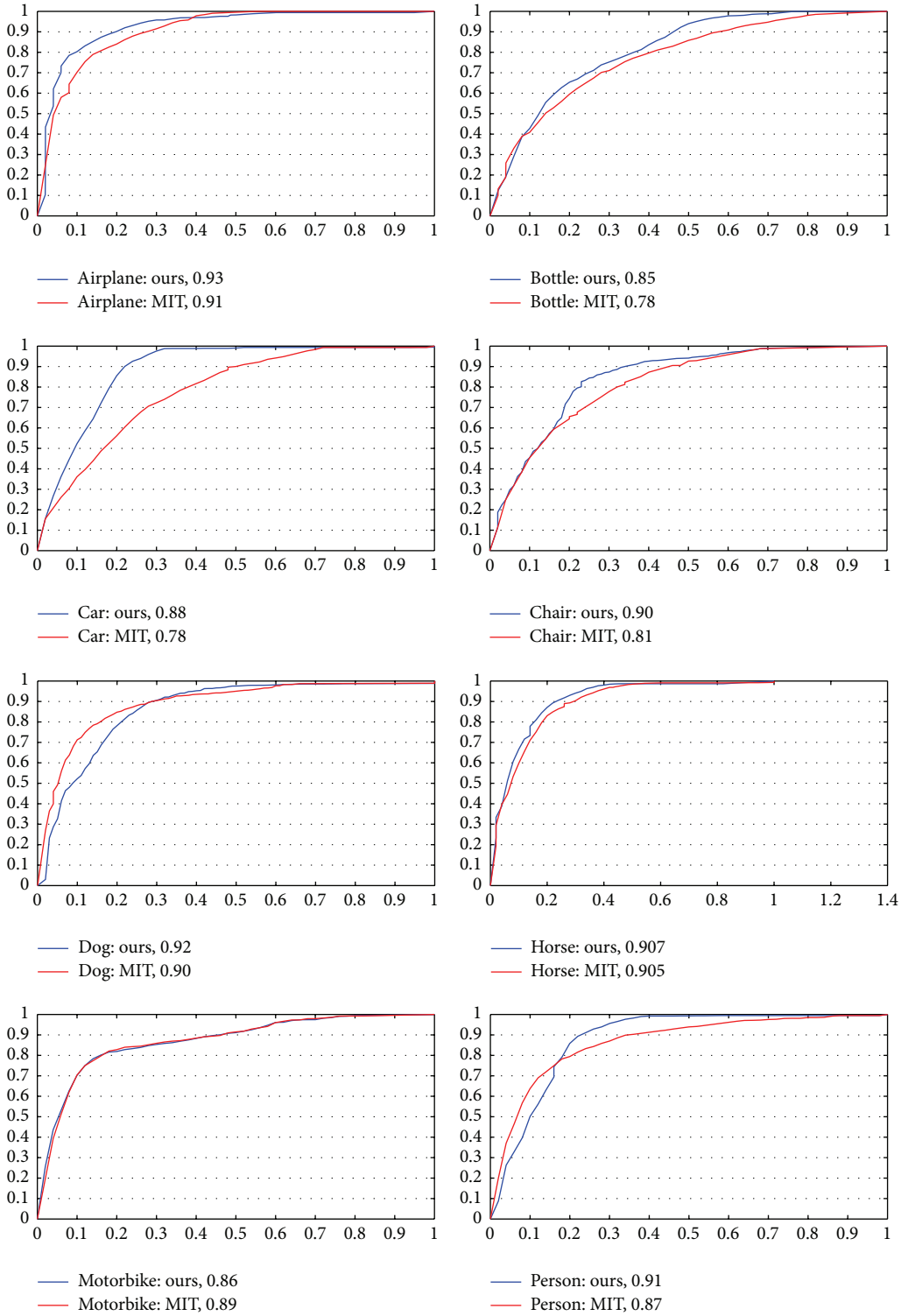


FIGURE 8: The figure shows the Receiver Operating Characteristic (ROC) for our model and MIT model. For each picture, the false alarm rate, on the x-axis, and the detection rate, on the y-axis. Besides, for each category, we calculate the averaging AUC scores of all the predictions, which are shown in above picture.

subjects might not have enough time to find the target objects and also promote the weight of center bias. On the other hand, if the viewing duration is too long, as the viewing proceeded, top-down or other factors (e.g., subjects feel bored and tired) come into play and gaze points become noisier. In addition, if the viewing duration is too long, gaze points may become the free-viewing.

Daily human activities involve a preponderance of visually guided actions, requiring observers to determine the presence and location of particular objects. Based on it, we researched how consistent human gaze points are across an image. Previous research and experience have shown that the gaze point location of several humans is strongly indicative of where a new subject will look, whether target-absent, and target-present. We implemented computational model for target-present in visual search and evaluated how well the model predicted subject's gaze points locations. In our experience, when subjects looked at a scene with a particular task, they consistently payed greater attention to the location of target objects and ignored the other saliency objects, such as text and people. So, our model combined the location of target as the high-level features. Ultimately, the model of attentional guidance predicted 95% of human agreement with the location of target object component providing the most explanatory power.

In this work we make the following contributions. We develop a collection of eye tracking data from the 11 people across 1307 images and have made it public for research use. It is the largest eye tracking database based on the visual search, which provides not only the accurate subjects' gaze points but also segmentation of target object for each image. In this search task, the location of target object is a dominating factor. We use machine learning to train a bottom-up, top-down model of saliency based on low-level, high-level, and center prior features. Finally, to demonstrate performance of our model, the same method was used to train MIT model.

For future work we may be interested in researching that the subjects' gaze points are tightly clustered in very small and specific regions, but our model selects a much more general region containing many objects without gaze points. We believe that the features of target object such as size, scale, and shape will lead subjects to fixate on target, which should be researched more carefully.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

The work is supported by NSF of China (nos. 61117115 and 61201319), the Fundamental Research Funds for the Central Universities, and NWPU "Soaring Star" and "New Talent and Direction" Program.

References

[1] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognitionwith visual attention," http://arxiv.org/abs/1412.7755.

[2] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models—their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[3] A. D. Hwang, H.-C. Wang, and M. Pomplun, "Semantic guidance of eye movements in real-world scenes," *Vision Research*, vol. 51, no. 10, pp. 1192–1205, 2011.

[4] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 185–198, 2010.

[5] W. S. Geisler and J. S. Perry, "Real-time foveated multiresolution system for low-bandwidth video communication," in *Human Vision and Electronic Imaging III*, vol. 3299 of *Proceedings of SPIE*, pp. 294–305, 1998.

[6] K. Shubina and J. K. Tsotsos, "Visual search for an object in a 3D environment using a mobile robot," *Computer Vision & Image Understanding*, vol. 114, no. 5, pp. 535–547, 2010.

[7] C. Siagian and L. Itti, "Biologically inspired mobile robot vision localization," *IEEE Transactions on Robotics*, vol. 25, no. 4, pp. 861–873, 2009.

[8] M. Casares, S. Velipasalar, and A. Pinto, "Light-weight salient foreground detection for embedded smart cameras," *Computer Vision & Image Understanding*, vol. 114, no. 11, pp. 1223–1237, 2010.

[9] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185–207, 2013.

[10] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," *Neural Information Processing Systems*, vol. 20, pp. 241–248, 2007.

[11] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.-S. Chua, "An eye fixation database for saliency detection in images," in *Computer Vision—ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5–11, 2010, Proceedings, Part IV*, vol. 6314 of *Lecture Notes in Computer Science*, pp. 30–43, Springer, Berlin, Germany, 2010.

[12] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proceedings of the IEEE 12th International Conference on Computer Vision (ICCV '09)*, pp. 2106–2113, Kyoto, Japan, September 2009.

[13] I. Van Der Linde, U. Rajashekar, A. C. Bovik, and L. K. Cormack, "DOVES: a database of visual eye movements," *Spatial Vision*, vol. 22, no. 2, pp. 161–177, 2009.

[14] J. H. Van Hateren and A. Van der Schaaf, "Independent component filters of natural images compared with simple cells in primary visual cortex," *Proceedings of the Royal Society B: Biological Sciences*, vol. 265, no. 1394, pp. 359–366, 1998.

[15] K. A. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva, "Modelling search for people in 900 scenes: a combined source model of eye guidance," *Visual Cognition*, vol. 17, no. 6-7, pp. 945–978, 2009.

[16] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, no. 4, pp. 219–227, 1985.

[17] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[18] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention,"

IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 5, pp. 802–817, 2006.

[19] D. Reisfeld, H. Wolfson, and Y. Yeshurun, "Context-free attentional operators: the generalized symmetry transform," *International Journal of Computer Vision*, vol. 14, no. 2, pp. 119–130, 1995.

[20] G. Heidemann, "Focus-of-attention from local color symmetries," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 26, no. 7, pp. 817–830, 2004.

[21] G. Kootstra, A. Nederveen, and B. de Boer, "Paying attention to symmetry," in *Proceedings of the British Machine Vision Conference*, pp. 1115–1125, Leeds, UK, September 2008.

[22] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," *Journal of Vision*, vol. 13, no. 4, article 11, 2013.

[23] S. Marat, T. Ho Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué, "Modelling spatio-temporal saliency to predict gaze direction for short videos," *International Journal of Computer Vision*, vol. 82, no. 3, pp. 231–243, 2009.

[24] N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga, "Saliency estimation using a non-parametric low-level vision model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '11)*, pp. 433–440, IEEE, Providence, RI, USA, June 2011.

[25] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vision Research*, vol. 49, no. 10, pp. 1295–1306, 2009.

[26] L. Zhang, M. H. Tong, and G. W. Cottrell, "SUNDAy: saliency using natural statistics for dynamic analysis of scenes," in *Proceedings of the 31st Annual Cognitive Science Conference*, Amsterdam, Netherlands, 2009.

[27] N. D. B. Bruce and J. K. Tsotsos, "Spatiotemporal saliency: towards a hierarchical representation of visual saliency," in *Attention in Cognitive Systems*, vol. 5395, pp. 98–111, Springer, Berlin, Germany, 2009.

[28] T. Avraham and M. Lindenbaum, "Esaliency (extended saliency): meaningful attention using stochastic image modeling," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 32, no. 4, pp. 693–708, 2010.

[29] B. Schölkopf, J. Platt, and T. Hofmann, "Graph-based visual saliency," *Advances in Neural Information Processing Systems*, vol. 19, no. 2006, pp. 545–552, 2006.

[30] H. J. Seo and P. Milanfar, "Nonparametric bottom-up saliency detection by self-resemblance," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 45–52, Miami, Fla, USA, June 2009.

[31] T. Liu, Z. Yuan, J. Sun et al., "Learning to detect a salient object," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 353–367, 2011.

[32] X. Hou and L. Zhang, "Saliency detection: a spectral residual approach," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 1–8, IEEE, Minneapolis, Minn, USA, June 2007.

[33] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search," *Psychological Review*, vol. 113, no. 4, pp. 766–786, 2006.

[34] M. Cerf, E. P. Frady, and C. Koch, "Faces and text attract gaze independent of the task: experimental data and computer model," *Journal of Vision*, vol. 9, no. 12, pp. 74–76, 2009.

[35] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *Journal of Vision*, vol. 11, no. 3, article 9, pp. 1–15, 2011.

[36] A. Borji, M. N. Ahmadabadi, and B. N. Araabi, "Cost-sensitive learning of top-down modulation for attentional control," *Machine Vision and Applications*, vol. 22, no. 1, pp. 61–76, 2011.

[37] R. J. Peters and L. Itti, "Beyond bottom-up: incorporating task-dependent influences into a computational model of spatial attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 1–8, Minneapolis, Minn, USA, June 2007.

[38] F. Baluch and L. Itti, "Mechanisms of top-down attention," *Trends in Neuroscience*, vol. 34, no. 4, pp. 210–224, 2011.

[39] M. Pomplun, "Saccadic selectivity in complex visual search displays," *Vision Research*, vol. 46, no. 12, pp. 1886–1900, 2006.

[40] J. Zelinsky Gregory, W. Zhang, B. Yu, X. Chen, and D. Samaras, "The role of top-down and bottom-up processes in guiding eye movements during visual search," in *Proceedings of the 19th Annual Conference on Neural Information Processing Systems (NIPS '05)*, vol. 18 of *Advances in Neural Information Processing Systems*, pp. 1569–1576, MIT Press, Cambridge, Mass, USA, 2005.

[41] W. Jian and Z. Xinbo, "Analysis of eye gaze points based on visual search," in *Proceedings of the IEEE International Conference on Orange Technologies (ICOT '14)*, pp. 13–16, Xian, China, September 2014.

[42] M. Jiang, J. Xu, and Q. Zhao, "Saliency in crowd," in *Computer Vision—ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., vol. 8695 of *Lecture Notes in Computer Science*, pp. 17–32, 2014.

[43] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research*, vol. 40, no. 10–12, pp. 1489–1506, 2000.

[44] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: a library for large linear classification," *Journal of Machine Learning Research*, vol. 9, no. 12, pp. 1871–1874, 2008.

[45] N. D. B. Bruce and J. K. Tsotsos, "Saliency based on information maximization," in *Advances in Neural Information Processing Systems 18.3*, pp. 155–162, MIT Press, 2005.