Improving Saliency Models by Predicting Human Fixation Patches

Rachit Dubey[1], Akshat Dave[2], and Bernard Ghanem[1]

[1] King Abdullah University of Science and Technology, Saudi Arabia
[2] University of California San Diego, USA

Abstract. There is growing interest in studying the Human Visual System (HVS) to supplement and improve the performance of computer vision tasks. A major challenge for current visual saliency models is predicting saliency in cluttered scenes (i.e. high false positive rate). In this paper, we propose a fixation patch detector that predicts image patches that contain human fixations with high probability. Our proposed model detects sparse fixation patches with an accuracy of 84% and eliminates non-fixation patches with an accuracy of 84% demonstrating that low-level image features can indeed be used to short-list and identify human fixation patches. We then show how these detected fixation patches can be used as saliency priors for popular saliency models, thus, reducing false positives while maintaining true positives. Extensive experimental results show that our proposed approach allows state-of-the-art saliency methods to achieve better prediction performance on benchmark datasets.

1 Introduction

Visual attention is an integral function of the Human Visual System (HVS). By focusing on a limited set of locations in the field of vision, the HVS prioritizes the distribution of perceptual resources to various locations in the visual field, making them more salient than others. There is substantial evidence of this non-uniform distribution in eye-fixation data [1], i.e. of parsimonious fixation on certain visual regions. Given a stimulus image, fixation points generally occupy a very small percentage of the overall number of image pixels. The thrifty allocation of processing resources is a result of *visual attention*. This process makes the recognition of patterns in the visual input computationally feasible. The phenomenon of visual attention in the HVS has received consideration in recent decades, particularly in the field of psychology and neuroscience [2]. These studies help understand the HVS better and are useful for a myriad of vision tasks including object recognition [3, 4], object detection [5] and action recognition [6].

Since eye fixations play an important role in object recognition/detection, a signicant amount of work has been done to automatically detect such fixations in images. Interest point detectors [7] output a particular set of points that are considered "interesting" and are extensively used in many computer vision systems. However, recent work [8] suggests that these detectors may have low perceptual relevance and are very weakly correlated to the HVS. In contrast to

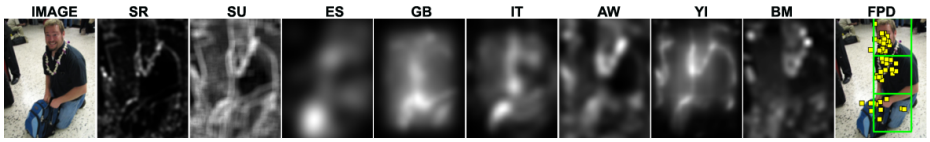

Fig. 1: **Elements in the background distract saliency models.** Example image showing saliency methods performing poorly due to background clutter. The last image shows human fixations in yellow and the detections of our proposed model in green. Our model is not significantly affected by background clutter, so it can be used to reduce false positives in various saliency methods.

interest point detectors, computational saliency models create pixel-level probability maps with the goal of predicting locations that have a high chance of attracting human attention. A wide variety of models have been proposed to compute visual saliency and have been shown to be useful in several vision tasks, such as object recognition [9] and image thumbnailing [10] among others.

Most saliency models generally perform well when applied to images containing a few salient regions. However, as recently reported in [11], a major challenge for these methods is predicting human attention in scenes containing various objects and distractors. Figure 1 shows an example of such a case wherein popularly used saliency methods tend to get distracted by background clutter resulting in poor prediction performance.

In this work, we propose a system that addresses this issue and improves the performance of popularly used saliency methods by reducing false positives. Our proposed system learns directly from human fixation data and utilizes biologically plausible low-level features to automatically and reliably identify image patches where humans might fixate in a free-viewing setting. We then use the detected patches as saliency priors to improve the performance of several state-of-the-art saliency models [12–20]. Through extensive experiments, we demonstrate the effectiveness of our method in improving the performance of these models on benchmark eye-tracking datasets. The models used in our experiments include some of the most recent and top performing saliency models in current literature.

Related Work: Following the classical algorithm of Itti and Koch [15], a number of researchers have worked to study visual saliency and the mechanism of eye fixations in the HVS. Itti and Baldi [21] studied bayesian surprise quantitatively to measure the extent to which humans direct their gaze to surprising items. A spectral residual approach for saliency detection was described in [12]. [13] proposed a bayesian framework for saliency using natural statistics. Harel et al. [14] proposed a bottom-up graph based saliency model. Judd et al. [22] employed machine learning techniques to develop a saliency model based on low, mid and high-level image features. More recently, a saliency method based on forward whitening of low-level features was proposed [16, 17] and been shown to outperform various other methods across several datasets in [11, 23]. We refer the reader to [24] for a recent survey on visual saliency modelling.

Most of the above described methods aim to predict the exact locations of human fixations yet it is unclear whether human fixations are deterministically

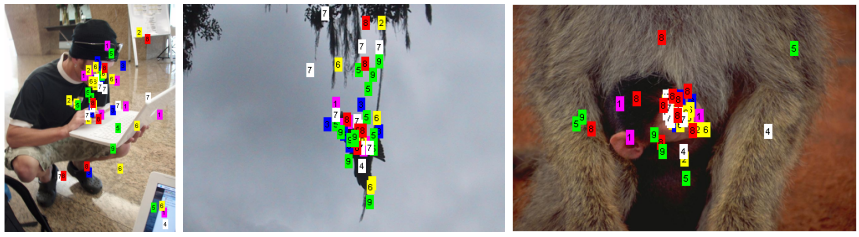

Fig. 2: **Humans tend to look at the same region but not the same locations.** Sample images with fixation points from different human observers plotted in different colors. Note the varying spatial distribution of fixations between observers.

repeatable locations in the visual field. As evident from data in [22], empirical evidence points to the conclusion that no two humans fixate on the *same* points while freely viewing the same image (refer to Figure 2 for an example). Nonetheless, these fixation points tend to lie in close proximity to each other and fixations from different humans do share similar spatial neighborhoods in the same image, thus, comprising an image region that we coin a *fixation region*. For simplicity and computational convenience, this spatial neighborhood can simply be modeled as a rectangular image patch.

As the main contribution of our work, we train a discriminative model to automatically identify *fixation patches* in an image. We then use the fixation patches as priors to reduce false positives in an effort to bridge the gap between current saliency models and human performance. This follows from our previous argument that saliency models tend to perform poorly on cluttered scenes.

The paper is organized as follows. Section 2 describes the feature extraction and training stages of our patch detector. Section 3 describes our proposed approach in the context of saliency model improvement. Experimental results are reported in Section 4 followed by discussion and analysis in Section 5.

2 Proposed Method: Fixation Patch Detector (FPD)

In this section, we give a detailed description of our learning based approach, which is is illustrated in Figure 3. Given an input image, we first divide the image into a set of non-overlapping patches. Patch extraction is followed by feature extraction wherein the extracted patches are represented by low-level and biologically inspired image features. Finally, a classifier is trained to identify image patches that attract human attention. These regions are denoted as *fixation patches*. In the rest of the paper, we refer to our proposed method as the Fixation Patch Detector (FPD). Next, we provide a detailed description of the steps involved in learning the FPD, namely feature extraction, and classifier learning.

2.1 Feature Representation

Although it is well known that human attention (even in the free-viewing case) is driven by both low-level (e.g. image intensity and gradients) and high-level

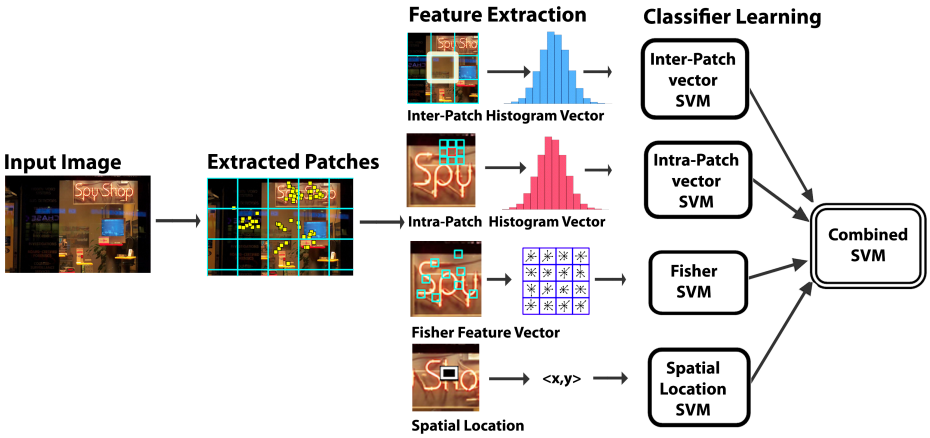

Fig. 3: **Overall framework for our proposed fixation patch detector.** (1) patch extraction, (2) feature extraction and (3) classifier learning.

features (e.g. familiar objects in the image), we focus on the former in this paper for simplicity. Our approach sheds light on how low-level features might influence human visual perception and discernibility. In this work, we propose a model that allows the prediction of image patches that have a high probability of attracting human fixations. We develop this model with the following three observations in mind.

(1) A viewer will fixate on a patch of the current image if he has fixated upon a similar patch before.
(2) A viewer will fixate on a patch that differs significantly from the patches in its local neighbourhood.
(3) A viewer will fixate on a patch within which there is a high degree of dissimilarity among pixels.

The basis for **(1)** is "familiarity", whereby humans are notoriously well equipped to recognize familiar objects, i.e. objects similar to those seen before. As such, we extend findings from previous studies [25, 26] and propose that humans also recognize and fixate on familiar salient regions in an image, i.e. previously seen salient patches. In other words, a human, who has fixated on a salient patch in one image, tends to fixate on a similar looking patch in another image. The basis for **(2)** is "surprise". As suggested by [21, 27], humans tend to fixate upon surprising items within the context of a scene. As such, patches in an image that significantly differ from their surroundings are expected to cause visual surprise in human viewers. Finally, observation **(3)** indicates that the content of an image patch plays a major role in attracting visual attention to this patch. In fact, humans tend to fixate *more* on image patches with heterogenous content than patches with a low pixel dissimilarity (i.e. intra-patch dissimilarity) [28, 29]. As such, the basis for **(3)** is "variance".

The above observations encourage the use of different sets of low-level image features. Therefore, we represent an image patch using four feature vectors. To

encode observation **(1)**, we use the popular Fisher kernel framework to extract a Fisher feature vector that describes local appearance information within a patch by relating it to previously encountered fixation patches. Observation **(2)** motivates the creation of an inter-patch self dissimilarity histogram, which describes how dissimilar an image patch is from its direct surroundings. Finally, we encode observation **(3)** using an intra-patch dissimilarity histogram, which describes how heterogenous a patch's appearance content is. To supplement these three features and to encode spatial information in our model, we use the normalized location of the patch center in the image to encode the spatial distribution of salient fixation patches. This spatial feature is motivated by the study of regional focus in the HVS, which is deemed independent of image properties, as well as, the work presented in [22, 30], which indicate that eye tracking datasets have a strong bias towards human fixations near the center of the image (the so-called *center bias*). As such, we divide each image $\mathbf{I}$ into a set of patches $P_{\mathbf{I}}$, each of which is represented by the four features described above and use them to train and test a fixation patch detector (FPD). This learned FPD will ultimately predict the likelihood of a patch attracting human visual attention.

Fisher Feature (F): Based on observation **(1)**, we represent each patch $\rho \in P_{\mathbf{I}}$ according to how similar it is to fixation and non-fixation patches observed in the training images. We describe how these ground truth patches are detected during training in Section 2.2. In this work, we choose to represent ρ using the Fisher kernel framework [31]. Here, a set of pixels in ρ (e.g. pixels with largest gradient energy) is selected as representative instances of a patch of an image. These instances are described by the conventional SIFT descriptor (spatially localized histograms of oriented gradients). To account for the variability of these SIFT vectors in the training set, a universal Gaussian Mixture Model (GMM) is constructed on the set of all training patches [32]. Using this universal GMM, the Fisher kernel feature of ρ is computed. If a patch $\rho \in \mathbf{I}$ has a set of representative instances Ω, its corresponding Fisher vector ($\mathbf{F}$) is computed as in Eq. (1).

$$\mathbf{F}(\rho) = \frac{1}{|\Omega|} \sum_{j \in \Omega} \nabla_{\gamma} \ln p(j|\gamma) \tag{1}$$

Here, ∇_{γ} is the gradient operator with respect to the Gaussian model parameter set γ (i.e. mean vector and variances). The Fisher feature is selected for its convenience in representing patches with different sized Ω using the same sized feature vector. In fact, our use of this type of feature is motivated by the noteworthy performance of Fisher features in other vision tasks including image classification [33]. Note that this Fisher-GMM framework is used in [34] for saliency detection, where it is assumed that images sharing global visual appearance are likely to share similar salient regions. Inspired by this work, we extend the framework on the principle that salient patches themselves are sampled from similar distributions that govern their local appearance. In that sense, our proposed method is a patch-based detector of human fixation.

Inter-Patch Dissimilarity Feature ($\mathbf{D_1}$): Motivated by observation **(2)**, we measure the dissimilarity of a patch $\rho \in P_{\mathbf{I}}$ with its direct neighbors. This dissimilarity can be computed in several ways, since it is highly dependent the features used to describe each of the patches. In this paper, we define the dissimilarity of ρ to its neighbors as the average difference in internal heterogeneity of the patches. In this case, a patch that is reasonably homogenous (i.e. whose appearance has minimal variation) is quite dissimilar from a patch that is heterogenous (i.e. whose appearance has significant variation). Therefore, patch ρ is represented by a self-dissimilarity descriptor, which is computed from the self-similarity descriptor in [35]. This self-dissimilarity descriptor is a matrix that is measured densely throughout the patch and serves as a measure of how heterogenous the patch's interior is. One of the main purposes for using self-similarity is that it captures the internal layout of a patch efficiently by unifying color, texture and edge patterns. It is noteworthy to mention that self-similarity has been shown to be an effective feature in several vision tasks and has been recently made computationally efficient for large resolution images [36]. To the best of our knowledge, this work is the first to apply self-similarity to the study of human fixations.

To compute the inter-patch dissimilarity feature $\mathbf{D}_1$ for ρ, we compute the histogram of the pair-wise Euclidean distances between the self-dissimilarity descriptor of ρ to those of its neighbours. This is the second ingredient of our FPD and it encodes "surprise".

Intra-Patch Dissimilarity Feature ($\mathbf{D_2}$): Based on observation **(3)**, we represent the inner content of patch ρ using a self-dissimilarity descriptor, as described before. The intra-patch dissimilarity feature $\mathbf{D}_2$ is then computed by constructing a histogram of the self-dissimilarity descriptors within patch ρ. This is the third patch feature and it encodes "variance".

Spatial Location Feature (C): Since the saliency of a patch might also be affected by its spatial location in the image, we represent each patch with a fourth feature $\mathbf{C}$, which is simply the normalized 2D coordinates of its center. This feature encodes "locality".

2.2 Classifier Training

This section describes how we train the classifier that will be used in FPD. Specifically, we discuss the details of preparing a fixation patch training dataset and subsequently the patch classification method.

Training Set Preparation: To train a fixation patch classifier, we require ground truth samples (positive and negative) of fixation patches. These patch samples are not readily available and manually annotating them is quite tedious, so we propose an automated way of inferring them using a dataset D_{all} of images and their corresponding human fixations from multiple observers. First, each image $\mathbf{I} \in D_{all}$ is divided into a set of non-overlapping square patches P_I,

where the size of each patch is taken to be $M \times M$ pixels. We score each patch in $P_{\mathbf{I}}$ based on the probability that a human fixation falls inside it. To compute this score, we construct an RBF kernel density estimate of the spatial distribution of all human fixation locations in image $\mathbf{I}$, denoted by $p_F(\mathbf{x}|\mathbf{I})$ where $\mathbf{x}$ is an individual pixel in $\mathbf{I}$. The score of patch $\rho \in P_{\mathbf{I}}$, denoted as $r(\rho)$, is computed as the pixelwise average probability of all pixels in ρ. Mathematically, we have $r(\rho) = \frac{1}{|\rho|} \sum_{\mathbf{x} \in \rho} p_F(\mathbf{x}|\mathbf{I})$. After scoring all patches in $\mathbf{I}$, we define ground truth fixation patches (labelled +1) as those with a score greater than a predefined threshold τ. Patches with scores less than τ are defined as non-fixation patches (labelled -1). Selecting a suitable τ for each image is not trivial. In our experiments, we take τ to be a predefined multiple of the peak value of $p_F(\mathbf{x}|\mathbf{I})$. Examples of $p_F(\mathbf{x}|\mathbf{I})$ and ground truth fixation patches are shown in Figure 4.

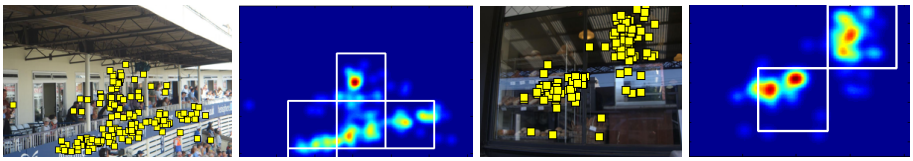

Fig. 4: **Ground truth fixation patches.** In each image, human fixations are plotted in *yellow*. The spatial density estimate of these fixations and the ground truth fixation patches (in *white*) are also shown. Non-fixation patches are patches whose average probability of containing a human fixation is below $\tau = 8\%$ of the peak probability.

Patch Classification: We apply PCA on the training set to reduce the dimensionality of the Fisher feature vector $\mathbf{F}$. Then, a standard RBF-SVM is trained on each of the four feature vectors separately: $\mathbf{F}$, $\mathbf{D}_1$, $\mathbf{D}_2$ and $\mathbf{C}$. Another SVM is then trained on the confidence values generated by the four individual SVMs. Learning is performed in this manner to *isolate* the effect of each feature type individually. This serves as the patch classifier for the FPD.

An important property of this fixation patch training set is that it is strongly unbalanced towards negative patches, i.e. non-fixation patches. This is primarily due to the sparsity of human fixations in an image. To overcome this significant data bias, conventional undersampling is incorporated in learning the classifier. In Figure 5, we show detection results using each of the individual SVMs as well as the combined classifier. Clearly, the latter classifier is able to effectively combine the individual SVM responses to accurately predict the fixation patches.

3 Saliency Enhancement

The patches predicted as non-fixations by our FPD method suggest that they are of little value to human observers. Therefore, giving less importance to such regions in saliency maps generated by popular visual saliency methods may improve the performance of these methods. Consider a set $\{s_k\}_{k=1}^{\Sigma}$, where s_k is

Fig. 5: **Patch Classification.** Detection results of SVMs using individual features (left to right): Fisher $\mathbf{F}$, inter-patch dissimilary $\mathbf{D}_1$, intra-patch dissimilarity $\mathbf{D}_2$, and location $\mathbf{C}$. The rightmost is the result of their combination and the proposed FPD

the k^{th} saliency model in a set Σ. Given an input image $\mathbf{I}$, the k^{th} model produces a saliency map $s_k(\mathbf{I})$. In an effort to improve saliency map prediction, we update $s_k(\mathbf{I})$ by combining the pixelwise saliency values with the corresponding pixelwise decision values returned by the FPD. This process effectively reduces the saliency of pixels inside predicted non-fixation patches. An added advantage of this strategy is that our FPD can also help to increase the saliency of pixels within fixation patches in case a saliency method misses strongly salient locations within these patch. In fact, we will show empirical evidence that verifies that our FPD allows popular saliency models to achieve better prediction performance on benchmark fixation datasets by reducing false positives, while maintaining true positives.

4 Experiments and Results

This section provides a quantitative analysis of our proposed FPD and empirical evidence showing that combining FPD with state-of-the-art saliency models improves their overall performance.

4.1 Dataset Description

To train the FPD, we use the MIT dataset [22] due to the diverse topical content of its images. This dataset contains 1003 high resolution images with human fixation data from 15 users per image. Maintaining the aspect ratio, each image is downsampled to no more than 2^{16} pixels. These images form our dataset D_{all}, which is randomly divided into D_{train} (903 images) and D_{test} (100 images) using K-fold cross validation (K = 10).

4.2 Parameter Settings

We use the same parameter settings in all our experiments. Based on cross validation results, each patch is taken to be $M \times M$ pixels with $M = 64$. Note that the FPD can be scaled to any patch size to suit application needs with adherence to runtime constraints. Of course, our proposed approach can be easily extended to multiple scales for more fine grained detection, accompanied by a linear increase in runtime. The value of the ground truth labeling threshold τ is set to 8% of the peak probability estimated using kernel density estimation.

4.3 FPD Performance Evaluation

Before evaluating the performance of the proposed FPD for saliency improvement, the predicted positive patches are compared to the ground-truth. The performance of the model is represented by its average positive accuracy p and average negative accuracy rate n. Using 10-fold cross validation, the FPD performs significantly well with $p = 84\%$ and $n = 84\%$ showing that the learned FPD can indeed be used to short-list and identify human fixation patches.

To evaluate the FPD's ability to enhance popular saliency methods, we compare the accuracy of each saliency method with and without our prior. We show results on D_{all} for 8 recent and state-of-the-art saliency models, namely **SR** [12], **SU** [13], **ES** [18], **GB** [14], **IT** [1] [15], **AW** [16, 17], **YI** [19] and **BM** [20].

Out of the many evaluation measures that have been used for comparing saliency models, ROC is the most widely used. The ROC curve is a measure of how well a saliency map can distinguish fixation and non-fixation points for different binary saliency thresholds. However, recent studies suggest that ROC is not always an ideal metric for comparison [37, 38, 11], since it only depends on the ordering of the fixations. As shown/argued in [37, 38], as long as the true positive rate is high, the area under the ROC curve (AUC) is always high regardless of the false positive rate. On the other hand, the main aim of our framework is to reduce false positives generated by a particular saliency model, while keeping this method's hit rate the same. Since ROC is affected more by the hit rate rather than false alarm rate, it is not suitable for a comprehensive evaluation of our proposed framework. Instead, we use two other popular evaluation measures, namely the Linear Correlation Coefficient (CC) [24, 11] and Normalized Scanpath Saliency (NSS) [39]. CC measures the strength of the linear relationship between a saliency map and the ground truth map, with an absolute value close to 1 indicating an almost perfectly linear relationship between the two. The NSS measures the average distance between the fixation saliency and zero with a larger NSS implying a greater correspondence between fixation locations and the saliency predictions [37].

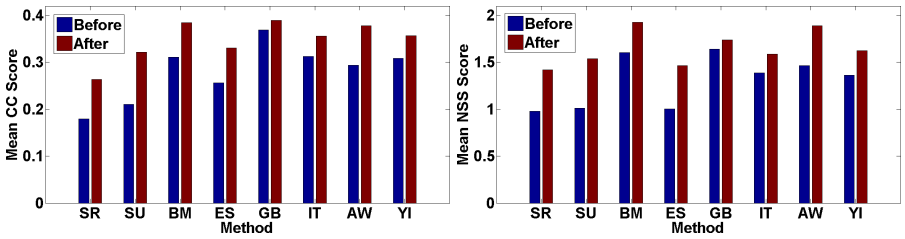

Fig. 6: **Performance evaluation on MIT dataset.** Average CC and NSS scores before and after applying the FPD on 8 state-of-the art saliency models. The FPD improves each score significantly for all tested methods.

[1] Note that there are different versions of Itti's model. Here, we used the best performing version in [14].

Figure 6 shows the average CC and NSS scores across all images in D_{all} for each saliency method before and after applying the FPD to the resulting saliency map.[2] Clearly, there is a significant improvement in both CC and NSS scores, which verifies the ability of the FPD to benefit each saliency model. Interestingly, our saliency prior is even able to enhance the performance of **BM** and **AW**, the top performing saliency models according to current literature [11, 23, 20].

Table 1: **Percentage of images for which CC and NSS is increased.** For each saliency method, we report the percentage of images for which the scores were increased after using the FPD.

Saliency Method	**SR**	**SU**	**BM**	**ES**	**GB**	**IT**	**AW**	**YI**
CC	81%	82%	81%	73%	67%	71%	76%	72%
NSS	86%	85%	80%	82%	67%	71%	80%	76%

Table 1 reports the percentage of images for which the CC and NSS scores were increased for each method. On average, we increase the CC score for over 75% of images and increase the NSS score for over 78% of images across all methods. In particular, our proposed FPD increases both scores for **SR**, **SU** and **BM** on over 80% of the images. Table 1 along with Figure 6 clearly show that higher performance is achieved when any of the saliency methods is simply combined with our proposed FPD.

Next, we test the performance of our FPD on another popular image dataset widely used for saliency evaluation. We use the FPD trained on D_{train} from the MIT dataset (refer to Section 4.1) and test it on the dataset in [40]. Since the training and test sets are different, this evaluates the generality of our method and its ability to transcend dataset specificities. Figure 7 shows mean scores across all images for each saliency method before and after using our method. The FPD is successful in improving the performance of all the saliency methods for this dataset as well. Interestingly, we once again observe a significant increase in performance for the top performing methods (**AW** and **BM**). Since combining these two models with our FPD significantly outperforms all other models, we recommend using this combination for future applications.

4.4 Center Bias

Even though the location feature in our model does not introduce any explicit center bias to the saliency maps (as our method focuses mainly on decreasing the false positives), it is still important to verify that the improvement in saliency performance reported in the previous section is not merely due to the location feature. It has often been pointed out in the literature that the issue of center bias is a major challenge for comparing saliency models. Several solutions have been proposed to eliminate center-bias effects, with the shuffled AUC metric being the

[2] Due to differences in image resolution, the reported scores of the saliency models are slightly different from those reported in [11, 23]. However, the relative performance of the models is not significantly affected.

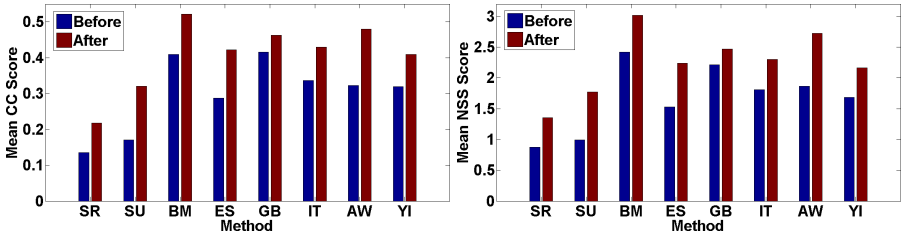

Fig. 7: **Performance evaluation on Bruce-Tsotsos dataset.** Average CC and NSS scores before and after applying the FPD on 8 state-of-the art saliency methods on the Bruce-Tsotsos dataset [40]. The FPD improves each score significantly for all tested methods for this dataset as well

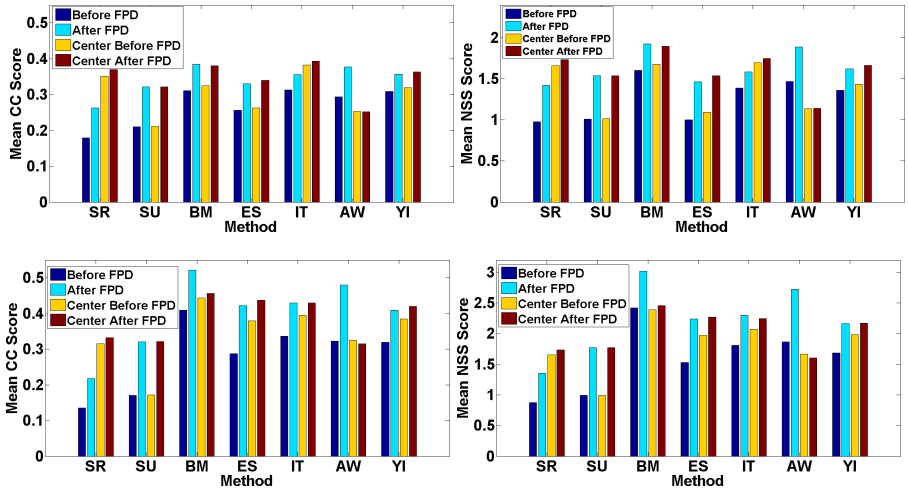

Fig. 8: **Comparison of FPD with explicit center bias.** Top row - performance of the FPD on center-biased saliency maps on the MIT dataset [22]. Bottom row - FPD performance on the Bruce-Tsotsos dataset [40].

most popularly used solution [13, 20, 23]. However, as discussed in Section 4.3, AUC is not an ideal metric for evaluating FPD. Moreover, shuffled AUC comes with the drawback of de-emphasizing genuine human fixations around the image center. For a more comprehensive evaluation, we explicitly introduce center bias to all the saliency methods as suggested in [11] by applying Gaussian blobs of varying sizes to each saliency map(centered at the image center and by increasing the size of the Gaussian by varying the sigma parameter, σ from 10 to 30 units). Our FPD is then applied to these center-biased saliency maps. This helps us evaluate the effectiveness of the FPD when the effect of the location feature is nullified due to the explicit addition of center bias. Since **GB** inherently adds a center bias and since we already showed an improvement in its performance, we exclude **GB** from this comparison.

Figure 8 shows the average CC and NSS scores after adding the center-bias and the performance of the FPD on both the original and center-biased

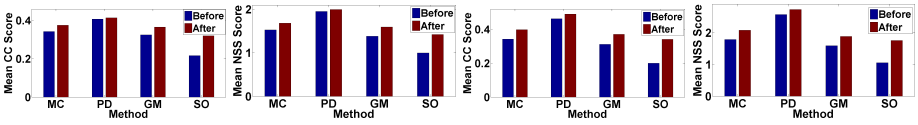

Fig. 9: **Performance evaluation of the FPD on salient object methods.** Average CC and NSS scores before and after applying the FPD on 4 salient object detectors on the MIT (first two plots) and Bruce-Tsotsos (last two plots) datasets.

saliency maps for both the MIT and Bruce-Tsotsos datasets when $\sigma = 20$(as in the previous section, no training was performed on the Bruce-Tsotsos dataset). As expected, adding explicit center bias increases both the CC and NSS scores for most of the saliency methods. However, it is encouraging to see that the performance of the FPD on the original saliency maps significantly outperforms the center-biased saliency maps for the top performing models, e.g. **BM**, **AW**, and **YI**. In addition, for the majority of the methods, the FPD increases the performance of the center-biased saliency maps as well. This is due to the fact that the center bias merely increases the saliency at the center. Our method, on the other hand, reduces the saliency of pixels within non-fixation patches, which could exist around the image center as well. A similar trend was also observed for all the different Gaussian sizes, i.e. when the σ value was 10 and 30 with the FPD increasing the performance in these cases as well. These results strongly suggest that the increase in the performance of saliency methods in the previous section is *not* simply due to the spatial feature **C** and that our proposed method plays a genuine role in enhancing the saliency methods themselves.

4.5 Improvement of Salient Object Detection Methods

Motivated by the work in [41] that showed that objects are highly effective in predicting human fixations, we also test our FPD's performance with 4 state-of-art salient object detectors, namely **SO** [42], **MC** [43], **PD** [44], and **GM** [45]. Figure 9 shows the performance of these methods before and after applying the FPD for fixation prediction on the MIT and Bruce-Tsotsos datasets. Consistent with our previous findings, the FPD improves the performance of the salient object methods as well.

5 Discussion

The experimental results in Section 4.3 show that the FPD is able to short-list a set of fixation patches in an image with high accuracy and a reasonably low false positive rate. We also compare of the performance of our FPD against FPDs built using the 8 different saliency methods as probability maps(in a strategy similar to that in Section 2.2). The accuracy of our proposed FPD was greater than 12% with respect to p accuracy and 14% with respect to n accuracy on average from these FPDs.

To summarize the results of Section 4.3, the performance of our FPD demonstrates significant promise. Not only does it improve the performance of state-of-the-art saliency models, but it does so with high consistency and high overall

magnitude. In particular, our method registers a significant improvement for top performing saliency methods **AW** and **BM**. This suggests that our proposed FPD can be used as a post processing step with saliency methods for pertinent vision applications. Figure 10 presents some qualitative results of our proposed approach. We show the output of the saliency maps generated by the eight different models, before and after applying FPD. The FPD benefits the saliency models by reducing the saliency of irrelevant background clutter, while maintaining the saliency of locations with high perceptual value.

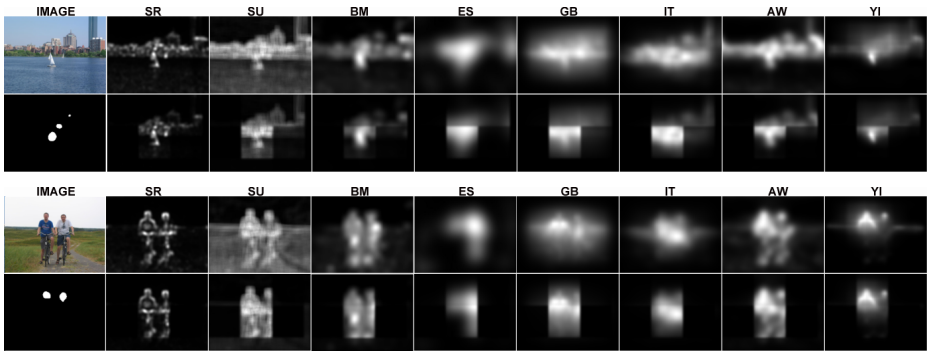

Fig. 10: **Qualitative results of our proposed approach.** Top and third row (left to right): Shows the input image and the saliency maps generated by 8 methods before applying the FPD. Second and last row (left to right): Shows the ground truth followed by our proposed FPD's saliency maps for each of the methods in the upper row.

Since humans fixate on objects of interest, the patches returned by the FPD have the potential of covering objects and being useful in higher level tasks. With this is in mind, we apply the FPD on the PASCAL VOC 2007 object detection dataset [46], which consists of 5011 images. Since ground-truth bounding boxes are provided for objects in all images in the dataset, we can evaluate how likely our predicted fixation patches overlap with these objects. In our experiments, we train on the entire MIT dataset D_{all} and use the VOC 2007 Trainval dataset (containing 5011 images) as the test set. We evaluate the performance of the FPD by first taking the union of all ground truth object detection windows in each image $\mathbf{I}$ of the test set to generate mask $\mathbf{M}_1$. We then take the union of all the salient fixation patches predicted by the FPD in $\mathbf{I}$ to form mask $\mathbf{M}_2$. Next, we find the intersection of all the predicted fixation patches with all the object detection windows, i.e. compute $|\mathbf{M}_1 \cap \mathbf{M}_2|$, and divide the intersection by the size of $\mathbf{M}_2$ to obtain the overall percentage overlap, i.e. compute $\frac{|\mathbf{M}_1 \cap \mathbf{M}_2|}{|\mathbf{M}_2|}$.

We repeat this process for all images in the test set and obtain an average overlap percentage of 59.6%. This implies that the predicted fixation patches are not only perceptually relevant in general but they also overlap reasonably well with object windows. For each fixation patch, we also calculate the highest

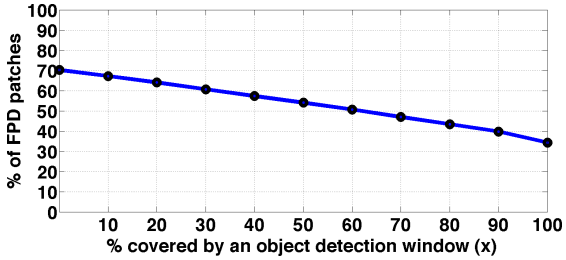

Fig. 11: **Experiments on the PASCAL VOC dataset.** Figure showing the percentage of patches that are covered by an object bounding box by more than x%. An object bounding box tends to have 60% or more coverage with roughly half the fixation patches showing the potential of the detected patches to capture salient object parts.

percentage of the patch that is covered by a single object window. This tells us the relevance of each fixation patch in covering a single object part. Figure 11 plots the percentage of predicted fixation patches that are overlapped by more than x% by an object window. For $x = 50$, this percentage is approximately 55%. This signifies that more than half of the predicted fixation patches are significantly covered by an object window. Interestingly, roughly 35% of these fixation patches overlap with the object windows completely and reside inside an object window. Since our patch size is much smaller than the average object window size, this could possibly suggest that FPD patches have the *potential* to find object parts.

6 Conclusion

In this paper, we propose a fixation patch detector (FPD) that predicts image regions where human fixations reside. We use these fixation patches as saliency priors to reduce the saliency of pixels within non-fixation patches leading to a consistent improvement in the prediction performance of several state-of-the-art saliency methods. The FPD significantly improves the performance of the top performing saliency methods, thereby suggesting that it can be used as a post processing step with state-of-the-art methods for future vision applications like object detection, image thumbnailing, etc.

Acknowledgement: Research reported in this publication was supported by competitive research funding from King Abdullah University of Science and Technology (KAUST).

References

1. Ross, J., Burr, D., Morrone, C.: Suppression of the magnocellular pathway during saccades. (Behavioural Brain Research)
2. Itti, L., Koch, C.: Computational modelling of visual attention. Nature Reviews Neuroscience **2** (2001) 194–203

3. Rutishauser, U., Walther, D., Koch, C., Perona, P.: Is bottom-up attention useful for object recognition. CVPR (2004)
4. Walther, D., Itti, L., Riesenhuber, M., Poggio, T., Koch, C.: Attentional selection for object recognition - a gentle way. BMCV (2002)
5. Endres, I., Hoiem, D.: Category independent object proposals. ECCV (2010)
6. Shapovalova, N., Raptis, M., Sigal, L., Mori, G.: Action is in the eye of the beholder: Eye-gaze driven model for spatio-temporal action localization. In: NIPS. (2013)
7. Mikolajczyk, Krystian; Schmid, C.: Scale & affine invariant interest point detectors. International journal of computer vision **60** (2004) 63–86
8. Dave, A., Dubey, R., Ghanem, B.: Do humans fixate on interest points? ICPR (2012)
9. Yang, L., Zheng, N., Yang, J., Chen, M., Chen, H.: A biased sampling strategy for object categorization. CVPR (2009)
10. Marchesotti, L., Cifarelli, C., Csurka, G.: A framework for visual saliency detection with applications to image thumbnailing. ICCV (2009)
11. Borji, A., Sihite, D., Itti, L.: Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. IEEE Transactions on Image Processing **22** (2013) 55–69
12. Hou, X., Zhang, L.: Saliency detection: A spectral residual approach. CVPR (2007)
13. Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.W.: Sun: A bayesian framework for saliency using natural statistics. Journal of Vision **8** (2008)
14. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. NIPS (2007)
15. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **20** (1998) 1254–1259
16. Garcia-Diaz, A., Fdez-Vidal, X.R., Pardo, X.M., Dosil, R.: Saliency from hierarchical adaptation through decorrelation and variance normalization. Image and Vision Computing **30** (2012) 51–64
17. Garcia-Diaz, A., Leborán, V., Fdez-Vidal, X.R., Pardo, X.M.: On the relationship between optical variability, visual saliency, and eye fixations: A computational approach. Journal of Vision **12** (2012)
18. Avraham, T., Lindenbaum, M.: Esaliency (extended saliency): Meaningful attention using stochastic image modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence **32** (2010) 693–708
19. Li, Y., Zhou, Y., Yan, J., Niu, Z., Yang, J.: Visual saliency based on conditional entropy. ACCV (2010)
20. Zhang, J., Stan, S.: Saliency detection: A boolean map approach. In: ICCV. (2013)
21. Itti, L., Baldi, P.: Bayesian surprise attracts human attention. NIPS (2006)
22. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. ICCV (2009)
23. Borji, A., Tavakoli, H., Sihite, D., Itti, L.: Analysis of scores, datasets, and models in visual saliency prediction. In: ICCV. (2013)
24. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence **35** (2012) 185–207
25. Soto, D., Humphreys, G.W., Heinke, D.: Working memory can guide pop-out search. Vision research **46** (2006) 1010–1018
26. Sheinberg, D.L., Logothetis, N.K.: Noticing familiar objects in real world scenes: the role of temporal cortical neurons in natural vision. The Journal of Neuroscience **21** (2001) 1340–1350
27. Yang, Y., Song, M., Li, N., Bu, J., Chen, C.: What is the chance of happening: a new way to predict where people look. ECCV (2010)

28. Poirier, F.J., Gosselin, F., Arguin, M.: Perceptive fields of saliency. Journal of vision **8** (2008) 14
29. Scharfenberger, C., Wong, A., Fergani, K., Zelek, J.S., Clausi, D.A.: Statistical textural distinctiveness for salient region detection in natural images. In: CVPR. (2013)
30. Le Meur, O., Le Callet, P., Barba, D.: Predicting visual fixations on video based on low-level visual features. Vision research **47** (2007) 2483–2498
31. Jaakkola, T., Haussler, D.: Exploiting generative models in discriminative classifiers. NIPS (1998)
32. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society, Series B (1977) 1–38
33. Perronnin, F., Snchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. ECCV (2010)
34. Marchesotti, L., Cifarelli, C., Csurka, G.: A framework for visual saliency detection with applications to image thumbnailing. ICCV (2009)
35. Shechtman, E., Irani, M.: Matching local self-similarities across images and videos. CVPR (2007)
36. Deselaers, T., Ferrari, V.: Global and efficient self-similarity for object classification and detection. CVPR (2010)
37. Zhao, Q., Koch, C.: Learning a saliency map using fixated locations in natural scenes. Journal of vision **11** (2011)
38. Judd, T., Durand, F., Torralba, A.: A benchmark of computational models of saliency to predict human fixations. Technical Report (2012)
39. Peters, R.J., Iyer, A., Itti, L., Koch, C.: Components of bottom-up gaze allocation in natural images. Vision research **45** (2005) 2397–2416
40. Bruce, N., Tsotsos, J.: Saliency based on information maximization. NIPS (2006)
41. Einhäuser, W., Spain, M., Perona, P.: Objects predict fixations better than early saliency. Journal of Vision **8** (2008) 18
42. Rahtu, E., Kannala, J., Salo, M., Heikkilä, J.: Segmenting salient objects from images and videos. ECCV (2010)
43. Jiang, B., Zhang, L., Lu, H., Yang, C., Yang, M.H.: Saliency detection via absorbing markov chain. In: ICCV. (2013)
44. Margolin, R., Tal, A., Zelnik-Manor, L.: What makes a patch distinct? In: CVPR. (2013)
45. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: CVPR. (2013)
46. Everingham, M., Van Gool, L., Williams, C., Winn, J., Zisserman, A.: The pascal visual object classes challenge 2007 (voc 2007) results (2007). In: URL http://www. pascal-network. org/challenges/VOC/voc2007/workshop/index. html. (2008)