Leveraging Human Fixations in Sparse Coding: Learning a Discriminative Dictionary for Saliency Prediction

(Invited Paper)

Ming Jiang
Department of Electrical and
Computer Engineering
National University of Singapore
mjiang@nus.edu.sg

Mingli Song
College of Computer Science
Zhejiang University
brooksong@ieee.org

Qi Zhao*
Department of Electrical and
Computer Engineering
National University of Singapore
eleqiz@nus.edu.sg

***Abstract*—This paper proposes to learn a discriminative dictionary for saliency detection. In addition to the conventional sparse coding mechanism that learns a representational dictionary of natural images for saliency prediction, this work uses supervised information from eye tracking experiments in training to enhance the discriminative power of the learned dictionary. Furthermore, we explicitly model saliency at multi-scale by formulating it as a multi-class problem, and a label consistency term is incorporated into the framework to encourage class (salient vs. non-salient) and scale consistency in the learned sparse codes. K-SVD is employed as the central computational module to efficiently obtain the optimal solution. Experiments demonstrate the superior performance of the proposed algorithm compared with the state-of-the-art in saliency prediction.**

Keywords*—Saliency, Visual Attention, Supervised Sparse Coding, Dictionary Learning, K-SVD

I. Introduction

Humans and other primates shift their gaze to allocate resources to the most relevant part of the visual world. This ability allows them to process the input data and react in real-time, and has evolutionary significance. In the computational domain, the same problem of information overload exists, and becomes the bottleneck of many artificial systems. Computational saliency models that predict important locations of a visual input have straightforward applications to a variety of real-world tasks such as target detection, video compression, and so on.

Over the years, a large body of computational saliency models have been proposed [1–11], most of which base themselves on low-level image features like color, intensity and orientation. A common limitation of these models is that they ignore higher-level semantic information of objects that also plays an important role - many times more important than low-level information - in directing gazes. The importance of high-level information in attention has been shown by several recent physiological [12, 13] and psychophysical [14, 15] experiments. The failure to encode high-level information in many existing models makes them perform less satisfactorily than the biological counterparts, especially in complex and semantically meaningful scenes. This has been referred to as the “semantic gap”, a long-lasting problem in the saliency literature.

In this paper, we aim to encode object-level semantics with supervised sparse coding, and reconstruct the sparse representations of an image into a saliency map. By collecting and utilizing positive and negative samples, it leverages human eye tracking data and constructs a dictionary with higher discriminative power. Furthermore, to encourage the class (salient vs. non-salient) and scale consistency of the learned codes, a label consistency term is explicitly incorporated in the proposed framework.

The contribution of the paper can be summarized as follows:

- First, we propose a novel algorithm of supervised dictionary learning with sparsity constraint to leverage eye tracking data for saliency prediction. As humans attend to semantic objects more than other regions, the algorithm automatically learns codes that represent semantically rich objects therefore bridging the “semantic gap”.
- Second, we explicitly model saliency prediction as a multi-class classification problem with two classes (salient and non-salient) and multiple scales. A label consistency term is used in the framework to enforce the consistency of the learned codes in the same class or scale in a principled manner.
- Third, besides the conventional center-surround based low-level features that have been proven to be effective in saliency prediction, we also use the Histograms of Oriented Gradients (HOG) that is widely used for object detection. By measuring patch statistics, it works complementarily with the center-surround features. More importantly, the effectiveness of HOG in object detection indicates its capability in encoding object-based saliency.

The proposed model has been compared with the state-of-the-art methods, especially those incorporating object de-

978-1-4799-0652-9/13/$31.00 ©2013 IEEE

* Corresponding author.

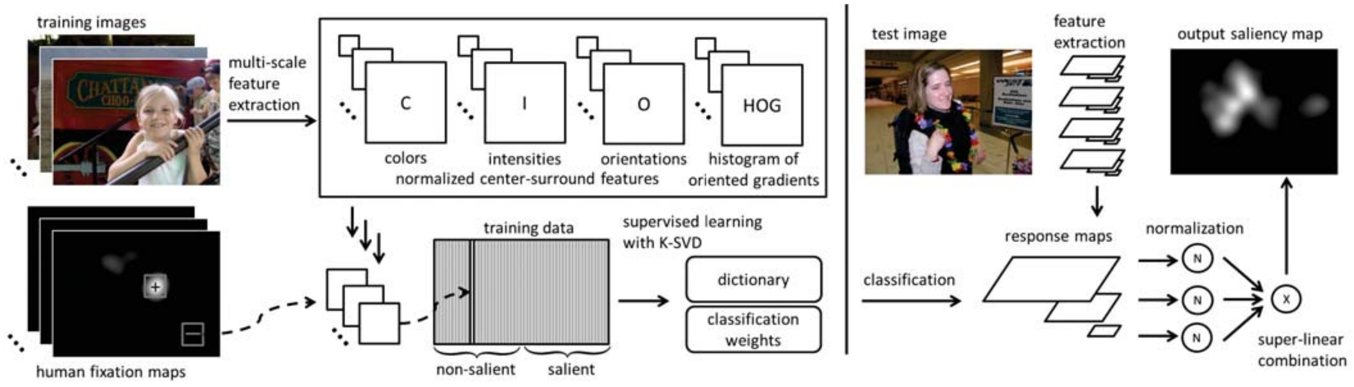

Fig. 1. An overview of the proposed saliency model. In the training phase, center-surround and HOG features are first extracted from a Gaussian pyramid of each training image. Then, from the ground-truth saliency map generated with human fixations, salient and non-salient image patches are sampled, whose features are later fed into a sparse coding algorithm to jointly learn a discriminative dictionary with basis functions as well as the classification weights. In the testing phase, the dictionary and weights are used to generate multi-scale saliency maps of a test image. These maps are finally normalized and super-linearly combined into the final saliency map.

tectors, over three public available eye tracking datasets. Our model achieves promising results without introducing any specific object detectors.

II. Related Works

A. Visual Saliency

Among the large body of saliency models, some of them are neutrally plausible, e.g., follow a structure rooted in the Feature Integration Theory (FIT) [16]. Back to the 1980s, Koch and Ullman [17] first proposed a computational saliency framework based on FIT, which Itti et al. [1] later implemented. Briefly, the model computes the center-surround feature maps in each of several low-level channels including color, intensity and orientation. These feature maps are then summed across multiple scales and normalized to yield a “conspicuity map” for each channel. Finally all the conspicuity maps are combined into a master saliency map by a linear or nonlinear integration to represent the probability that attention would be attracted by each image pixel. Along this line, a number of computational models [2–7] have been proposed.

Based mostly on low-level features, various computational algorithms were developed to infer saliency of different feature channels. Following a Bayesian framework, Zhang et al. [8] proposed the Saliency Using Natural statistics (SUN) model, which considered not only the features in a local neighborhood, but also the rarity of a feature compared to the global statistics of the current image. Harel et al. [9] proposed another probabilistic model called the Graph-Based Visual Saliency (GBVS). By constructing a grid graph over the image, the low-level feature maps are combined using a Markov chain based approach. The Attention based on Information Maximization (AIM) model was proposed by Bruce and Tsotsos [10], which was based on the information maximization theory. It samples RGB colors from image patches, and estimates a probability density function by maximizing the information that a region conveys relative to its surroundings. Recently, Hou et al. [11] proposed an image signature approach to detect spatially sparse foreground regions from complex background, using the inverse Discrete Cosine Transform (DCT) transform of the signs in the cosine spectrum.

To fill the “semantic gap” between computational saliency models and human performance, specifically-trained object detectors have been incorporated into saliency models. For example, faces have been shown to attract attention independent of tasks, and several recent models [18–20] combined face detection as a separate visual cue with traditional low-level features to improve saliency detection. Furthermore, Judd et al. [21] proposed a support vector machine (SVM) based learning approach to linearly combine face, pedestrian and car detectors with low- and mid-level features. To some extent, the integration of multiple object detectors increases the prediction performance, yet it is barely possible to scale such algorithms to the large number of object categories in real life. To approach this challenge, this paper leverages human data with supervised sparse coding and a set of features to effectively represent low-level as well as object-level information. The saliency maps learned directly from the human data are therefore capable of encoding semantic objects that are not limited to any specific categories.

B. Sparse Coding

Sparse coding has been effectively used in a variety of tasks in computer vision, such as face recognition [22], object detection [23], and has recently been applied to saliency prediction. Hou and Zhang [24] proposed the Incremental Coding Length (ICL) to measure the respective entropy gain of sparse features. By selecting features with large coding length increments, it achieves attention selectivity in both dynamic and static scenes. Borji and Itti [25] proposed a saliency framework that projects image patches to the feature space of a dictionary learned with sparse coding and measures local and global patch rarities in the feature space to detect salient image regions. In Guo and Chen’s work [26], two dictionaries - one from low-level image features and another from eye fixations - are jointly learned. They use dense SIFT as local features combined with global color and orientation probability to represent image patches sampled around fixation locations. The sparse coefficients of an image are applied to the fixation dictionary to reconstruct a saliency map. These efforts focus either on measuring the low-level statistics of natural images(as background) thus differentiating foreground from the modelled background, or trying to represent common features shared by the salient image regions. In this paper, we aim to integrate these two approaches by learning sparse object representations from both salient and non-salient regions, which allows us to build a discriminative model for saliency prediction.

III. Learning a Discriminative Dictionary for Saliency Prediction

This work provides a general framework for saliency prediction. In particular, we aim to learn out semantic object information to fill the “semantic gap”. Two distinctions from conventional object detection methods are that: a) salient objects detected by the proposed method are not restricted to a certain number of interested categories, and b) instead of using pre-defined image sets with object labels, the training data are sampled from images viewed by human subjects. Figure 1 shows an overview of the proposed saliency framework.

A. Feature Extraction and Sampling

Center-Surround Features. Following the conventional saliency model by Itti et al.[1], an input image is first subsampled into a Gaussian pyramid of S scales from $1/1$ (scale 0) to $1/256$ (scale 8). At each scale, the image is decomposed into seven feature channels, including the Red/Green (C_{RG}) and Blue/Yellow (C_{BY}) color contrast channels, the Intensity (I) channel, and four local Orientation ($O_\theta, \theta \in \{0°, 45°, 90°, 135°\}$) channels computed using Gabor filters. For each of these channels, center-surround feature maps are computed by subtracting every center pixel at a fine scale $c \in \{3, 4, 5, 6\}$ by the corresponding surround pixel at a coarse scale $s = c + \delta$ with $\delta \in \{2, 3\}$, yielding 8 center-surround maps in total. Finally the computed center-surround maps are normalized with operator $\mathcal{N}(\cdot)$. The local center-surround contrast between the center scale c and the surround scale s can be computed as:

$$\mathcal{L}_l(c, s) = \mathcal{N}(|\mathcal{F}_l(c) \ominus \mathcal{F}_l(s)|) \quad (1)$$

where $\mathcal{F}_l(t)$ represents the raw feature map of channel $l \in \{C_{RG}, C_{BY}, I, O_{0°}, O_{45°}, O_{90°}, O_{135°}\}$ at scale t, and the operator $\ominus$ denotes the pixel-by-pixel subtraction between the center and surround maps.

HOG Features. HOG features have been widely used in object detection [27], for its ability to capture object texture and contour information against noises or environmental changes. To encode image statistics as a complementary cue to the pixel-level center-surround features, locally normalized HOG representation with both contrast-sensitive and contrast-insensitive orientation bins is incorporated. We follow the construction in [28] to define a 31-dimensional dense representation of an image at each particular scale. For example, generating HOG features at scale $c+3$ can be done by dividing the image at scale c into 8×8 non-overlapping cells. For each cell, a histogram of 9 gradient orientations over the pixels is constructed and then normalized with respect to the gradient energy in a 2×2 neighborhood around it.

In our approach, a dictionary of image features is learned from both salient and non-salient image patches. First, a ground-truth saliency map of an image is derived from human eye tracking data. Particularly, to construct a fixation map from eye tracking data, each fixation location is represented as a white pixel (and non-fixated locations as black ones), and the fixation map is then blurred with a Gaussian kernel to generate the ground-truth saliency map. The intensities of the blurred saliency map indicate the fixation density at each particular image pixel. Second, to sample the salient patches for dictionary learning, we randomly extract p pixels from the top 30% salient regions and q pixels from the bottom 30% salient regions in the ground-truth saliency map. For every selected image pixel, we extract its $r \times r$ neighborhood in each scale and combine the center-surround and HOG features into a feature vector. Therefore, p positive samples and q negative samples are extracted at each scale.

B. Dictionary Learning with Class and Scale Consistency

Sparse coding has found support in the biological domain – sparsity is not only the response property of neurons in area V1, but also that of areas deeper in the cortical hierarchy [29]. To encode higher-level information of salient and non-salient image patches, we propose a sparse coding approach that uses a sparse linear combination of low-level features for an efficient representation of image features in relation to saliency. Basically, the learned dictionary contains representative features as basis functions to linearly reconstruct the training image patches with minimum error. Mathematically, let $Z = [z_1, \cdots, z_i, \cdots, z_M] \in R^{N \times M}$ be a set of N-dimensional image features extracted from labelled salient or non-salient image patches, and we aim to obtain discriminative sparse codes $X = [x_1, \cdots, x_i, \cdots, x_M] \in R^{K \times M}$ with a dictionary $D = [d_1, \cdots, d_k, \cdots, d_K] \in R^{N \times K}$. The objective of this dictionary learning problem can be formulated as:

$$< D, X > = \arg\min_{D,X} \|Z - DX\|_F^2 \quad s.t. \forall i, \|x_i\|_0 \leq T \quad (2)$$

where the $\|\cdot\|_F$ denotes the Frobenius norm and the term $\|Z - DX\|_F^2$ represents the reconstruction error. T is a sparsity constraint factor that stands for the maximum number of non-zero entries in each sparse code x_i. Saliency prediction is casted as a multi-class classification problem, where each class corresponds to a class label (i.e., salient or non-salient) and a scale, and we in this work follow the label consistent K-SVD (LC-KSVD) approach proposed by Jiang et al. [31], to learn a set of discriminative sparse codes and a linear classifier simultaneously for saliency prediction. Specifically, this is done by adding two regularization terms to Equation 2. One term enforces the discrimination capabilities for salient versus non-salient image patches at different scales, which encourages the input data sampled from the same class (salient or non-salient) and the same scale to have very similar sparse representations. The other is a classification error term, which allows the learned sparse codes to be predictive with a linear classifier. Accordingly, the objective of the saliency prediction problem can be formulated as follows:

$$\begin{aligned} < D, A, X, w > = \arg\min_{D,A,X,w} \|Z - DX\|_F^2 \\ +\alpha\|U - AX\|_F^2 + \beta\|v^T - w^T X\|_2^2 \quad s.t. \forall i, \|x_i\|_0 \leq T \end{aligned} \quad (3)$$

The three terms $\|Z - DX\|_F^2$, $\|U - AX\|_F^2$, and $\|v^T - w^T X\|_2^2$ represent the reconstruction error, the discriminative sparse-code error, and the linear classification error respectively. The coefficients α and β control the relative contributions of the corresponding terms.

Here the matrix $U = [u_1, \cdots, u_i, \cdots, u_M] \in \{0,1\}^{K\times M}$ are the discriminative sparse codes of input Z for classification. Each column u_i is a 'discriminative' sparse code corresponding to an input sample z_i. $A \in R^{K\times K}$ is a linear transformation matrix that transforms the original sparse codes in X to be most discriminative. To explain this in the saliency context, assume the input data $Z = (Z_0^1, \cdots, Z_0^S, Z_1^1, \cdots, Z_1^S)$ is a set of image features sampled at s scales, where Z_0^s and $Z_1^s, s = 1\ldots S$ respectively represent non-salient and salient sub-matrices of Z at scale s. The matrix U can be defined as:

$$U \equiv \begin{pmatrix} U_0^1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 & 0 & 0 \\ 0 & 0 & U_0^S & 0 & 0 & 0 \\ 0 & 0 & 0 & U_1^1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & U_1^S \end{pmatrix} \qquad (4)$$

where $U_l^s, l \in \{0,1\}, s = 1\ldots S$ are all matrices of ones. Thus, the term $\|U - AX\|_F^2$ enforces that the sparse codes X can approximate the discriminative sparse codes U with a linear transformation A.

In the classification error term $\|v^T - w^T X\|_2^2$, $w = [w_1, \cdots, w_k, \cdots, w_K]^T \in R^K$ is the classification weights to reconstruct the ground-truth saliency labels $v = [v_1, \cdots, v_i, \cdots, v_M]^T \in [0\cdots 1]^M$ with the sparse representations X. Note that instead of using binary labels for classification, v_i represents the ground-truth saliency value of the i-th input sample, which is the central intensity of the image patch in the ground-truth saliency map.

To find the optimal solution for all parameters simultaneously, Equation 3 can be rewritten as:

$$< D, A, X, w >= \arg\min_{D,A,X,w} \left\| \begin{pmatrix} Z \\ \sqrt{\alpha}U \\ \sqrt{\beta}v^T \end{pmatrix} - \begin{pmatrix} D \\ \sqrt{\alpha}A \\ \sqrt{\beta}w^T \end{pmatrix} X \right\|_F^2 \quad s.t.\forall i, \|x_i\|_0 \le T \qquad (5)$$

By letting $Z' = (Z^T, \sqrt{\alpha}U^T, \sqrt{\beta}v)^T$ and $D' = (D^T, \sqrt{\alpha}A^T, \sqrt{\beta}w)^T)$, Equation 5 can be formulated as

$$< D', X >= \arg\min_{D',X} \|Z' - D'X\|_F^2 \quad s.t.\forall i, \|x_i\|_0 \le T \qquad (6)$$

As a generalization of data clustering, the above dictionary learning problem can be efficiently solved by the K-SVD algorithm [32].

Figure 2 shows a visualization of HOG features in the dictionary learned from the FIFA dataset [18] that contains images of faces. The dictionary is divided into 8 blocks, each with 100 bases representing salient or non-salient patches at one of the four scales. The patch size is 7×7. As illustrated, our algorithm is able to extract salient object-level image features like eyes and noses (scale 3), faces and heads (scale 4, 5) and upper-body postures (scale 6). It is worth noting that quite a few learned non-salient features are long contours, junctions, corners and textures, where regions corresponding to them are easily misclassified as salient regions using conventional low-level feature based saliency models.

C. Saliency Prediction

The obtained dictionary $D = \{d_1, \cdots d_k, \cdots d_K\}$, transform parameters $A = \{a_1, \cdots a_k, \cdots a_K\}$ and classification

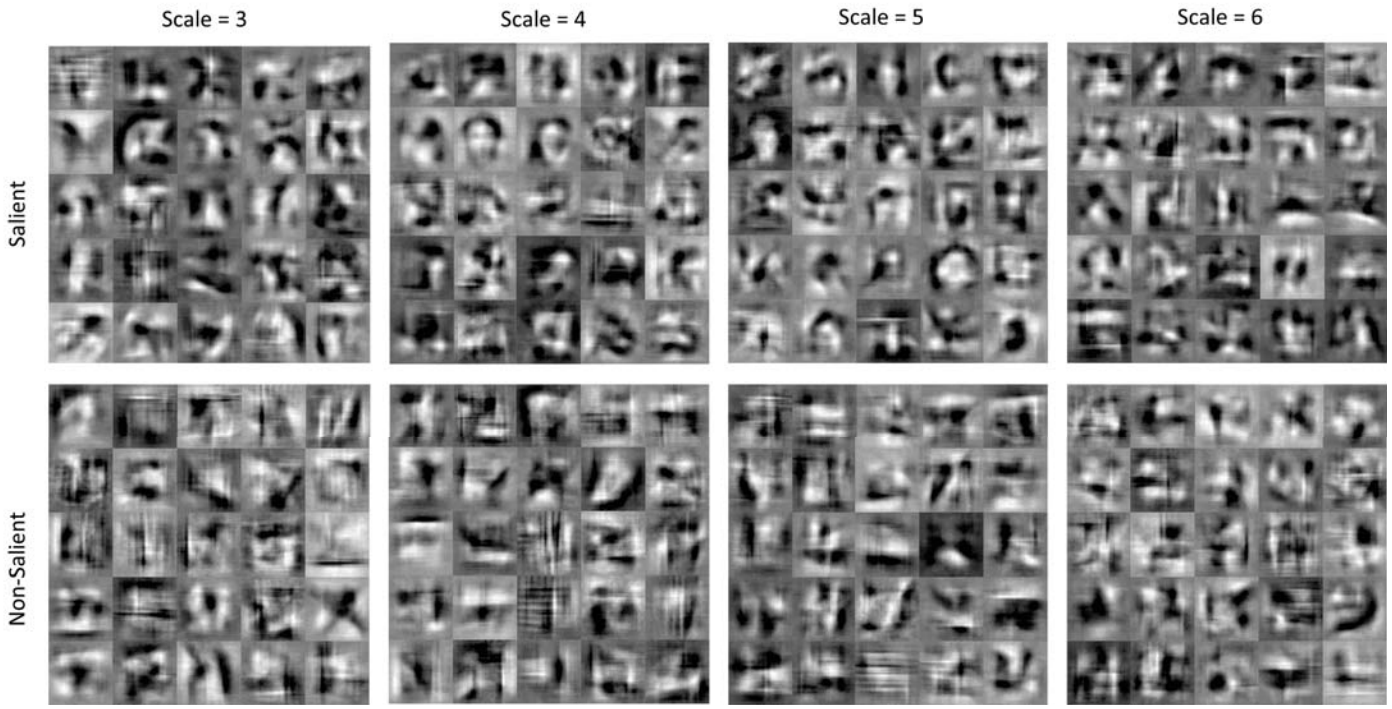

Fig. 2. The example of HOG features (inverted into image patches using the iHOG toolbox [30]) in the learned basis functions from the FIFA dataset [18].

weights $w = \{w_1, \cdots w_k, \cdots w_K\}^T$ in the above supervised training phase can be used to predict the saliency map of a new image. Note that D, A and w cannot be directly used for testing since they are jointly normalized in D' in the LC-KSVD algorithm, i.e. $\forall k, \|d_k^T, \sqrt{\alpha} a_k^T, \sqrt{\beta} w_k)^T\|_2 = 1$. Instead, given a test feature vector z, the sparse codes x and the predicted saliency value v can be computed as follows:

$$x = \arg\min_x \|z - \hat{D}x\|_2^2 \quad s.t. \|x\|_0 \le T \tag{7}$$

$$v = \exp \hat{w}^T x \tag{8}$$

where $\hat{D}$ and $\hat{w}$ are denoted as:

$$\begin{aligned} \hat{D} &= \{\frac{d_1}{\|d_1\|_2}, \cdots \frac{d_k}{\|d_k\|_2}, \cdots \frac{d_K}{\|d_K\|_2}\} \\ \hat{w} &= \{\frac{w_1}{\|d_1\|_2}, \cdots \frac{w_k}{\|d_k\|_2}, \cdots \frac{w_K}{\|d_K\|_2}\} \end{aligned} \tag{9}$$

For each scale of the features, we use a sliding window approach to compute the saliency value of every pixel to generate a saliency map. The saliency maps at all scales are then normalized and combined to generate the master saliency map. Empirically, we find that using a super linear combination instead of linearly summing up across all scales leads to better prediction performance and visualization results.

Algorithm 1 summarizes the proposed algorithm.

Algorithm 1 Learning a Discriminative Dictionary to Predict Saliency

```
// Training stage:
Input: A set of training images and corresponding eye fixations collected from human subjects viewing the images;
Output: Dictionary $D$, classification weights $w$;
1) Convolve a Gaussian kernel with the fixation map to generate a ground-truth saliency map for each image (Sec. III-A);
2) Compute multi-scale center-surround and HOG features (Sec. III-A);
3) Extract sample features from salient and non-salient image patches (Sec. III-A);
4) Learn the dictionary $D$ and weights $w$ by optimizing Equation 3 with K-SVD (Sec. III-B).

// Test stage:
Input: The dictionary $D$, classification weights $w$, a test image;
Output: Saliency map of the input test image;
1) Compute multi-scale center-surround and HOG features for the test image (Sec. III-A);
2) At each scale, find a $r \times r$ window around each pixel and extract features for the particular pixel (Sec. III-C);
3) Compute the saliency value at each pixel using the learned parameters (Sec. III-C);
4) Normalize the multi-scale saliency maps and combine them into a final saliency map (Sec. III-C).
```

IV. Experiments

This section reports experimental results to validate the proposed algorithm. Sec. IV-A introduces three datasets used in this work for comparative experiments. Sec. IV-B discusses metrics to evaluate the algorithms, and Sec. IV-C presents quantitative comparison results of the proposed work and state-of-the-art counterparts.

A. Datasets

The following three eye tracking datasets are used in this work, all of which contain a large number of objects:

The MIT dataset [21] contains 1003 landscape and portrait images collected from Flickr and LabelMe [33]. Image contents include indoor and outdoor environments and a wide variety of object categories like humans, animals, cars, text, etc. This dataset contains eye fixation data from 15 subjects free-viewing each of the images for 3 seconds..

The FIFA dataset [18] contains 181 colored natural images. The images contain faces in various poses, sizes, and positions, as well as text, cellphones, cars, fruits and toys, etc. Eye-tracking data were collected from 8 subjects with a 2 second free-viewing task.

The NUSEF dataset [34] contains scenes and objects with strong emotions. It contains 758 natural images, and each image was viewed by 25 subjects for 5 seconds.

All images in each dataset were randomly divided into a training set with 80% images and a test set with 20% images, for training and evaluation purposes respectively.

B. Evaluation Metrics

In the saliency literature, there are several widely used criteria to quantitatively evaluate the performance of saliency models by comparing the saliency prediction with eye movement data. One of the most common evaluation metrics is the area under the receiver operator characteristic (ROC) curve (i.e. AUC) [35]. A problem with this metric is that it is significantly affected by the center bias effect [36], so the shuffled AUC was then introduced [8] to address this problem. Particularly, to calculate the shuffled AUC, negative samples are selected from human fixational locations from all images in a dataset (except the test image), instead of uniformly sampling from all image locations.

In addition, the Normalized Scanpath Saliency (NSS) [37] and the Correlation Coefficient (CC) [38] are also used to measure the statistical relationship between the saliency prediction and the ground truth. NSS is defined as the average saliency value at the fixation locations in the normalized predicted saliency map which has zero mean and unit standard deviation, while the CC measures the linear correlation between the saliency map and the ground-truth map. The three measures are complementary and provide a more objective evaluation of the various models.

C. Performance Evaluation

We evaluate the performance of the proposed model, as well as six state-of-the-art saliency models that are public

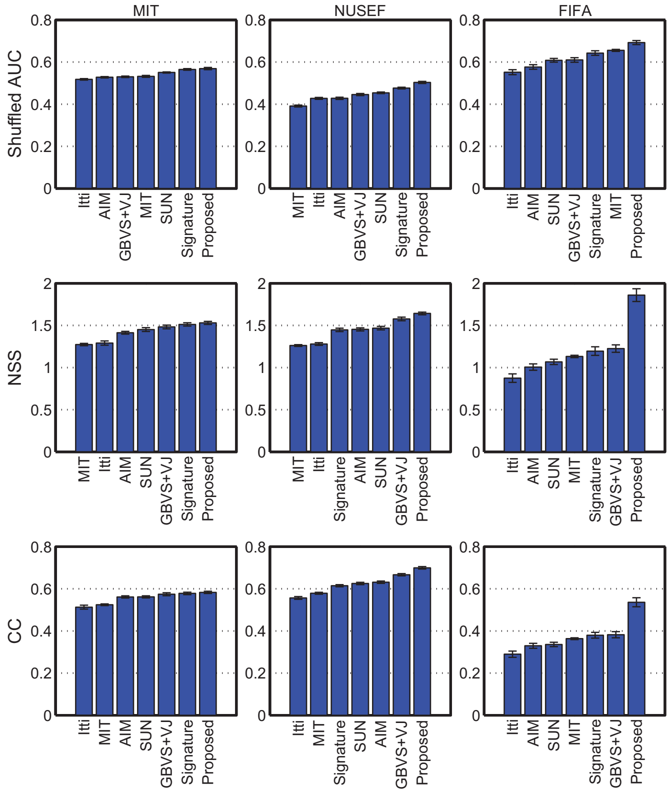

Fig. 3. Quantitative comparison of models. The prediction accuracy is measured with the shuffled AUC, NSS and CC scores. The bar values indicate the mean performance on 10 randomly chosen training and test sets. Error bars indicate the standard error of the mean.

available. Two of them are bottom-up models combined with object detectors (i.e. MIT [21] and GBVS+VJ [18], while the others are purely bottom-up models, including Itti et al.'s model implemented by Harel [39], the AIM [10], the SUN [8] and the Image Signature [11]. Both the MIT saliency model and ours are learning based, so the same training and test image sets are used in these two models. For all the three datasets, the dictionary size of the proposed model is 800. Two important parameters in the proposed saliency model are the coefficients α and β that determine the contributions of the label consistent regularization term and the classification error term in the K-SVD optimization. We noticed that the impacts of the parameters more or less depend on the dataset, largely due to the different semantic natures of these datasets. For example, the MIT dataset contains a wider variety of object categories as well as larger regions of low-level contrasts, while the FIFA dataset contains only a few semantically categorized objects, among which most are visually similar. In our experiments, we exhaustively tested the model performance with different α and β, and empirically set the parameters with the best performance for each dataset. Finally, it has been suggested that smoothing the final saliency map could increase the prediction performance [11]. Hence, for a fair comparison, all the saliency maps of each model are smoothed with the Gaussian kernel which leads to its best performance.

Figure 3 illustrates the comparative results. The proposed model generally outperforms other models over the three datasets. Although an unfair comparison, our model, without an explicit incorporation of any specifically-trained object detectors, performs better than those with object detectors. In fact, it learns out object related codes automatically and the learned object (parts) are not restricted to any pre-defined categories and scale well to the large number of categories in real life. It is worth noting that, the more object categories, the more data are needed to train an effective model. As shown in Figure 3 , our model performs significantly better on the FIFA dataset in which eye gazes are consistently drawn by only a few categories of object (mostly the faces), while in the MIT and NUSEF datasets that have a larger variety of object categories, performance would be further boosted with a larger amount of training data.

Figure 4 shows a qualitative comparison between our model and the state-of-the-art. As illustrated, by learning object-level codes from eye fixation data, the advantages of our model are twofold. First, positive samples from where most eye fixations occur are effectively used to detect the most salient objects. Learning to detect salient objects directly from eye tracking data makes our model more scalable than those explicitly adding object detectors. As seen in Figure 4, for example, not only human faces (rows 1, 2, 3, 5), but other objects like animal faces (rows 4, 9), text (rows 5, 8), or even salient objects in low-contrast image regions (row 10) can also be detected. Second, the negative samples from the background include considerable amount of regions with long edges, junctions, corners and textures, thus the proposed model effectively removes edge effects that lead to false saliency detection (rows 4, 6, 7, 10). We believe these advantages are mainly contributed by the incorporation of higher-level sparse representations, as well as the class and scale consistency of the proposed learning algorithm.

V. Conclusion

This paper presents a saliency model that uses supervised sparse coding to learn saliency maps from eye fixations. Object-level statistics are represented by combining HOG feature with local center-surround maps, and object-level semantics are automatically learned to effectively fill the "semantic gap" in saliency prediction. Experimental results on different datasets with various metrics consistently demonstrate that the proposed model outperforms the leading saliency algorithms, some of which explicitly incorporate object detectors.

Acknowledgment

This research was carried out at the SeSaMe Centre. It is supported by the Singapore NRF under its IRC@SG Funding Initiative and administered by the IDMPO.

References

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

[2] D. Parkhurst, K. Law, E. Niebur *et al.*, "Modeling the role of salience in the allocation of overt visual attention," *Vision research*, vol. 42, no. 1, pp. 107–124, 2002.

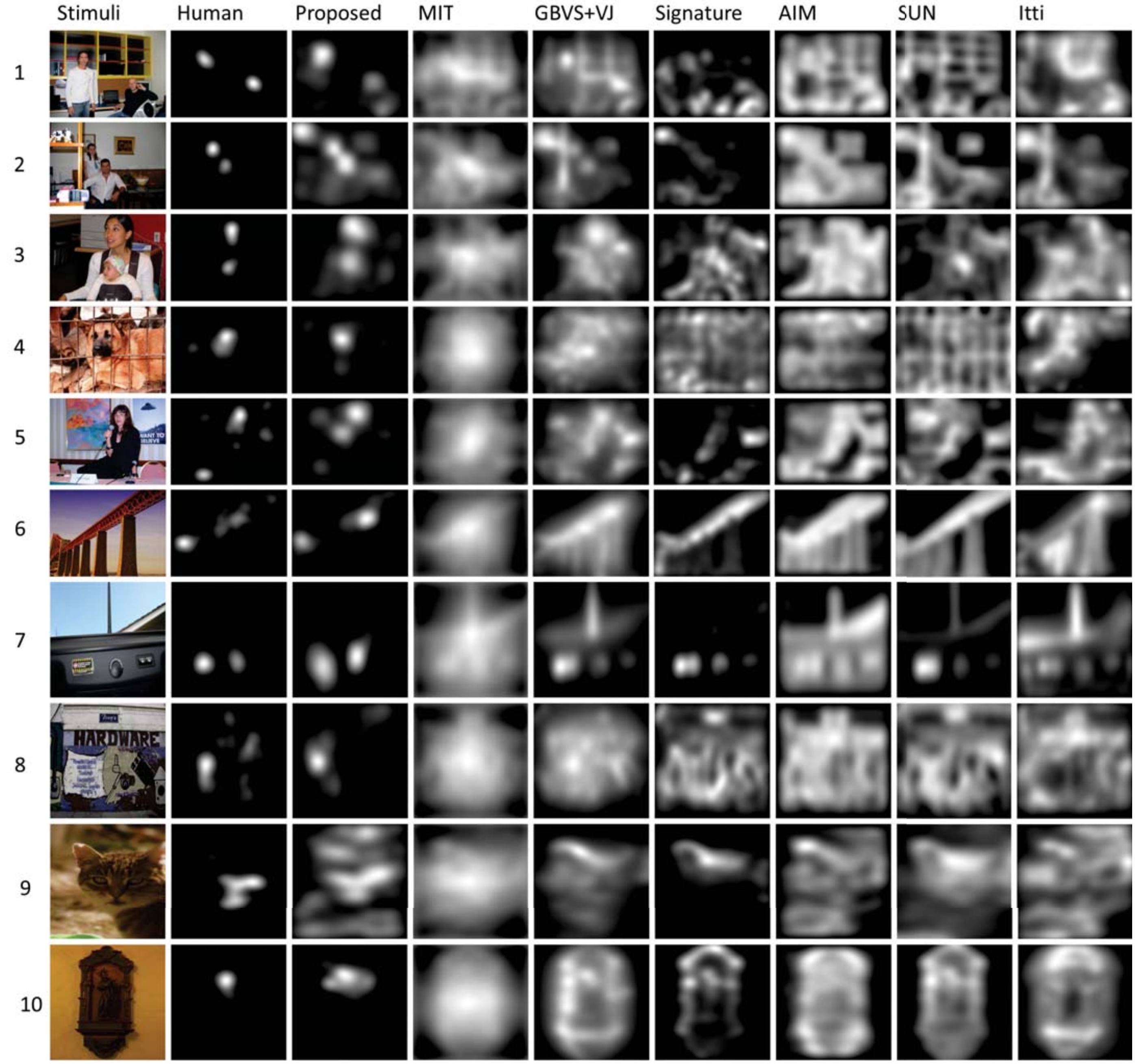

Fig. 4. Qualitative results of the proposed model and the state-of-the-art models over samples from FIFA $(1-3)$, NUSEF $(4-6)$ and MIT $(7-10)$ datasets.

[3] A. Oliva, A. Torralba, M. Castelhano, and J. Henderson, "Top-down control of visual attention in object detection," in *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, vol. 1. IEEE, 2003, pp. I–253.

[4] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[5] T. Foulsham and G. Underwood, "What can saliency models predict about eye movements? spatial and sequential aspects of fixations during encoding and recognition," *Journal of Vision*, vol. 8, no. 2, 2008.

[6] C. Masciocchi, S. Mihalas, D. Parkhurst, and E. Niebur, "Everyone knows what is interesting: Salient locations which should be fixated," *Journal of vision*, vol. 9, no. 11, 2009.

[7] S. Chikkerur, T. Serre, C. Tan, and T. Poggio, "What and where: a bayesian inference theory of attention," *Vision research*, vol. 50, no. 22, pp. 2233–2247, 2010.

[8] L. Zhang, M. Tong, T. Marks, H. Shan, and G. Cottrell, "Sun: A bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, 2008.

[9] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems*, 2007, pp. 545–552.

[10] N. Bruce and J. Tsotsos, "Saliency, attention, and visual

search: An information theoretic approach," *Journal of Vision*, vol. 9, no. 3, 2009.

[11] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, p. 194, 2012.

[12] E. Craft, H. Schütze, E. Niebur, and R. von der Heydt, "A neural model of figure–ground organization," *Journal of Neurophysiology*, vol. 97, no. 6, pp. 4310–4326, 2007.

[13] S. Mihalas, Y. Dong, R. von der Heydt, and E. Niebur, "Mechanisms of perceptual organization provide auto-zoom and auto-localization for attention to objects," *Journal of Vision*, vol. 10, no. 7, pp. 979–979, 2010.

[14] W. Einhäuser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency," *Journal of Vision*, vol. 8, no. 14, 2008.

[15] A. Nuthmann and J. Henderson, "Object-based attentional selection in scene viewing," *Journal of vision*, vol. 10, no. 8, 2010.

[16] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[17] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry." *Hum Neurobiol*, vol. 4, no. 4, pp. 219–27, 1985.

[18] M. Cerf, E. 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, 2009.

[19] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *Journal of vision*, vol. 11, no. 3, 2011.

[20] ——, "Learning visual saliency by combining feature maps in a nonlinear manner using adaboost," *Journal of Vision*, vol. 12, no. 6, 2012.

[21] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Computer Vision, 2009 IEEE 12th international conference on*. IEEE, 2009, pp. 2106–2113.

[22] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Robust sparse coding for face recognition," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 625–632.

[23] S. Agarwal and D. Roth, "Learning a sparse representation for object detection," in *Proceedings of the 7th European Conference on Computer Vision-Part IV*. Springer-Verlag, 2002, pp. 113–130.

[24] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," *Advances in neural information processing systems*, vol. 21, pp. 681–688, 2008.

[25] A. Borji and L. Itti, "Exploiting local and global patch rarities for saliency detection," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 478–485.

[26] K. Guo and H.-T. Chen, "Learning sparse dictionaries for saliency detection," in *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*. IEEE, 2012, pp. 1–5.

[27] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[28] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1627–1645, 2010.

[29] H. Lee, C. Ekanadham, and A. Ng, "Sparse deep belief net model for visual area v2," *Advances in neural information processing systems*, vol. 20, pp. 873–880, 2008.

[30] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba, "Inverting and visualizing features for object detection," *arXiv preprint arXiv:1212.2278*, 2012.

[31] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent k-svd," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1697–1704.

[32] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionries for sparse representation," *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4311–4322, 2006.

[33] B. Russell, A. Torralba, K. Murphy, and W. Freeman, "Labelme: a database and web-based tool for image annotation," *International journal of computer vision*, vol. 77, no. 1, pp. 157–173, 2008.

[34] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T. Chua, "An eye fixation database for saliency detection in images," *Computer Vision–ECCV 2010*, pp. 30–43, 2010.

[35] B. W. Tatler, R. J. Baddeley, I. D. Gilchrist *et al.*, "Visual correlates of fixation selection: Effects of scale and time," *Vision research*, vol. 45, no. 5, pp. 643–659, 2005.

[36] B. W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *Journal of Vision*, vol. 7, no. 14, 2007.

[37] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vision research*, vol. 45, no. 18, pp. 2397–2416, 2005.

[38] N. Ouerhani, R. von Wartburg, H. Hugli, and R. Muri, "Empirical validation of the saliency-based model of visual attention," *Electronic letters on computer vision and image analysis*, vol. 3, no. 1, pp. 13–24, 2004.

[39] J. Harel, "A saliency implementation in matlab," 2010.