Label Consistent Quadratic Surrogate Model for Visual Saliency Prediction

Yan Luo[1], Yongkang Wong[2], Qi Zhao[1*]

[1]Department of Electrical and Computer Engineering, National University of Singapore
[2]Interactive & Digital Media Institute, National University of Singapore
{luoyan, yongkang.wong, eleqiz}@nus.edu.sg

Abstract

*Recently, an increasing number of works have proposed to learn visual saliency by leveraging human fixations. However, the collection of human fixations is time consuming and the existing eye tracking datasets are generally small when compared with other domains. Thus, it contains a certain degree of dataset bias due to the large image variations (*e.g., *outdoor scenes* vs. *emotion-evoking images). In the learning based saliency prediction literature, most models are trained and evaluated within the same dataset and cross dataset validation is not yet a common practice. Instead of directly applying model learned from another dataset in cross dataset fashion, it is better to transfer the prior knowledge obtained from one dataset to improve the training and prediction on another. In addition, since new datasets are built and shared in the community from time to time, it would be good not to retrain the entire model when new data are added. To address these problems, we proposed a new learning based saliency model, namely Label Consistent Quadratic Surrogate algorithm, which employs an iterative online algorithm to learn a sparse dictionary with label consistent constraint. The advantages of the proposed model are three-folds:* **(1)** *the quadratic surrogate function guarantees convergence at each iteration,* **(2)** *the label consistent constraint enforces the predicted sparse code to be discriminative, and* **(3)** *the online properties enable the proposed algorithm to adapt existing model with new data without retraining. As shown in this work, the proposed saliency model achieves better performance than the state-of-the-art saliency models.*

1. Introduction

In the recent advances in sensor technology, computer vision systems are undoubtedly facing great difficulty to process the increasing number of pixels available from multiple visual sources. To tackle the information overload problem, visual saliency detection has emerged to be an efficient solution to detect the regions of interest to enhance existing computer vision system. For example, image and video compression [15], visual tracking [25] and object recognition [28, 2, 33].

Conventional saliency models employ a straightforward bottom-up solution to predict visual saliency [13, 14, 17, 39]. Recently, learning based saliency prediction models were proposed to leverage the power of machine learning techniques and human knowledge (from human fixation maps), and decipher the pattern to better predict the saliency regions. These models generally achieve stable performance on various datasets. However, the conventional learning based models in saliency prediction assume that the training data is fully observed and there exist sufficient training data. It is important to state that the collection of human fixations is time consuming and the existing eye tracking datasets are relatively small when compared with other domains. Thus, it consists a certain degree of dataset bias due to the large variations in images (*e.g.,* outdoor scenes *vs.* emotion-evoking images) and human subjects. In addition, existing learning based models are trained and evaluated within the same dataset where cross dataset validation is not yet a common practice. It is important to note that existing models are unable to adapt a learned model with new training data unless the model is retrained from scratch.

To address all the aforementioned problems, we propose a new saliency prediction model, namely *Label Consistent Quadratic Surrogate (LCQS) Algorithm*, which employs an iterative online dictionary learning framework with label consistent constraint. The novelty of the proposed model are as followed: First, we adapt the Quadratic Surrogate (QS) algorithm [26] to solve the sparse dictionary learning problem. It enables the dictionary learning process to depend on one training sample at a time, which provides good training efficiency and convergence rate. Second, we add label consistent constrain in the dictionary learning process

*Corresponding author.

to ensure that the learned sparse dictionary can generate discriminative sparse code for saliency prediction. Last but not least, the proposed saliency model can adapt a trained dictionary with new training data. This allows us to leverage the prior knowledge from other dataset to improve the quality of dictionary on a new dataset. This property also addresses the limitation in the number and size of available human fixations datasets. As shown in Section 5, the proposed saliency prediction model achieves better performance than the state-of-the-art saliency models.

The remaining of the paper is organized as follows. Section 2 describes the related work. Sections 3 and 4 elaborate our proposed online saliency framework with LCQS algorithm. Section 5 demonstrates qualitative and quantitative results, and Section 6 concludes the paper.

2. Related Works

2.1. Saliency model

Modeling visual attention has recently raised a great amount of research interest [4, 13, 14, 17, 18, 39]. The first saliency model was proposed by Koch and Ullman [23]. Based on [23], Itti *et al.* [17] proposed a bottom-up computational model with center-surround feature to detect conspicuous regions. Zhang *et al.* [39] exploited bottom-up saliency cues from natural statistics to measure the improbability of a local patch. Harel *et al.* [13] introduced a Graph-Based Visual Saliency (GBVS) model to weigh the dissimilarity between two arbitrary positions to detect the conspicuous regions. In [14], Hou *et al.* considered saliency detection as a figure-ground separation problem and employed sparse signal analysis to solve it. Recently, Zhang *et al.* [38] proposed a Boolean map based saliency model to compute the saliency map by analyzing the topological structure of Boolean maps.

The aforementioned bottom-up saliency models are straightforward solutions for saliency detection. Recently, learning based saliency prediction models were emerged to leverage the power of machine learning techniques and human knowledge (from human fixation maps), and decipher the pattern to better predict the saliency regions [18, 21, 40]. Jiang *et al.* [18] proposed a learning based model based on the Label Consistent K-SVD (LC-KSVD) algorithm [20], where the goal is to fill the semantic gap between computational saliency models and human behavior. Despite the results demonstrated superiority over non-learning based methods, this method has to be retrained from scratch is new training data are built and shared in the community.

2.2. Online Learning

Machine learning techniques are widely employed in signal processing, neuroscience and computer vision community. Most of the machine learning techniques employ a *batch learning* framework, where the model was trained once with a set of training data. The training process are generally slow and the quality of the trained model are confined by the quality of the training data. In contrast, online machine learning is a model of induction that learns one instance at a time. There are two unique advantages of online learning: **(1)** the training instances in each learning stage is very small, which results cheaper training cost and better model convergence, **(2)** it avoids model retraining from scratch and can adapts the existing model with new data in the future.

There exists a variety of online learning algorithms [6, 7, 9, 22, 29, 36]. The normal herd algorithm [7] was introduced to herd a Gaussian weight vector distribution by trading off velocity constraints with a loss function. In [36], the soft confidence-weighted algorithm was proposed to address the limitation of the confidence-weighted algorithm [6], which is prone to wrongly change the parameters of the distribution. Kivinen *et al.* [22] presented a kernel-based algorithm with Stochastic Gradient Descent (SGD) in an online setting. This method suffered from the high algorithmic complexity and extensive memory cost for large number of training instances. The aforementioned works depended on linear model and SGD method within the original feature space, which is not capable to fully decipher the complicated patterns. Based on K-SVD algorithm [1], Jiang *et al.* [19] proposed the LC-KSVD model to learn an overcomplete dictionary over a set of training instances, and enforce the learned dictionary to be more discriminative. However, this method do not satisfy the mathematical properties of online learning and required to retrain model when new training instances are available. In [20], Jiang *et al.*extend the LC-KSVD model with an incremental learning framework with SGD. However, there is no evidence to guarantee the convergence properties in each learning stage and the model does not support online learning with new training data. In this work, we adopt an online dictionary learning algorithm, namely Quadratic Surrogate (QS) algorithm [26], as the solution for learning the sparse dictionary.

3. Sparse Coding Based Saliency Model

3.1. Feature Extraction & Sampling

Itti *et al.* [16] has proved that center-surround feature is effective for the modeling of visual attention. Histogram of Oriented Gradients (HOG) [8] has been widely accepted as one of the best features to capture the edge or local shape information in detection. In this work, we adopt center-surround and HOG feature as input of the proposed model.

Center-Surround Feature. Following the conventional saliency model by Itti *et al.* [17], an input image is subsampled into a Gaussian pyramid of S scales from $1/1$

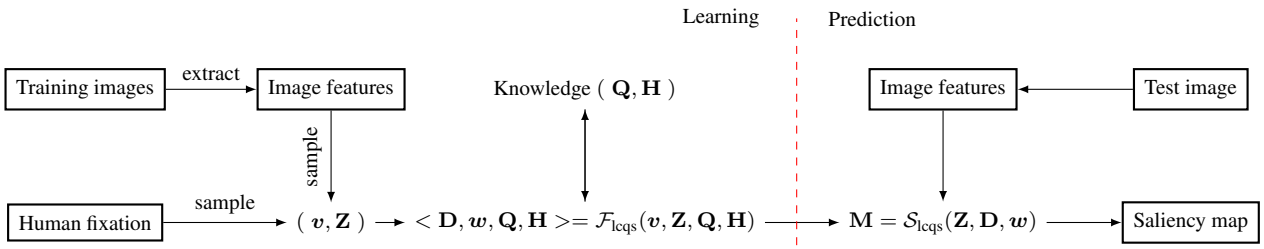

Figure 1: An overview of the LCQS saliency model.

(scale 0) to $1/256$ (scale 8). The image is decomposed into seven feature channels at each scale, including two color contrast channels (Red-Green C_{RG} and Blue-Yellow C_{BY}), intensity channel I and four local orientation channels ($O_\theta,\ \theta \in \{0°, 45°, 90°, 135°\}$) computed using Gabor filters. For each of these channels, center-surround feature maps are computed by subtracting each center pixel at a fine scale $c \in \{3, 4, 5\}$ by the corresponding surrounding pixels at a coarse scale $s = c + \delta,\ \delta \in \{2, 3\}$, yielding 6 center-surround maps in total.

Histogram of Oriented Gradients Feature. HOG feature can capture object's texture and contour information against noises or environmental changes. Locally normalized HOG representation with both contrast-sensitive and contrast-insensitive orientation bins is incorporated. Similar to the feature construction in [11], we define a dense representation of an image at each particular scale $\tilde{c}$ where $\tilde{c} = c + 3$.

Sampling Strategy. In this work, dictionary is learned from both salient and non-salient training samples. First, a ground truth saliency map, $\mathbf{M}_{GT}$, of an image is derived from visual fixation maps from human eye tracking data. Specifically, each fixation location is represented as a white pixel while non-fixated locations are represented with black pixels, followed by a blur operation with Gaussian kernel. Second, given the feature maps $\mathbb{F}$ from various scale c or $\tilde{c}$, training samples can be extracted based on the saliency value $v_{x,y}$ of corresponding location on $\mathbf{M}_{GT}$. Each training sample is represented as a duple $< v, \boldsymbol{z} >$ which consists of: **(1)** the saliency value v; and **(2)** feature vector $\boldsymbol{z}$ by extracting $r \times r$ neighborhood at corresponding pixel from $\mathbb{F}$ and concatenating all the center-surround and HOG features. We select $\bar{n}$ and $\bar{m}$ samples with the highest and lowest saliency values, respectively. In this work, we empirically set both value to 3600.

3.2. Label Consistent Quadratic Surrogate Model

In the context of sparse representation, the objective is to approximate a given sample as a linear combination of a small number of basis elements, where these basis elements form the subspaces of a feature space. This feature space is thought to be overcomplete such that any given sample can be represented with a relatively small set of basis elements. In this work, sparse coding approach is employed to learn an efficient representation of image features in relation to visual saliency in an online fashion. An overview of the proposed model is shown in Fig. 1. Under the formal mathematical formulation, let us suppose that $\mathbf{D} = [\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_k] \in \mathbb{R}^{m \times k}$ is a dictionary and each column $\boldsymbol{d}_i$ is known as a basis. Given a set of training feature samples, $\mathbf{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n] \in \mathbb{R}^{m \times n}$, extracted from salient and non-salient image patches, conventional sparse dictionary learning problem [1, 24, 26] is solved by optimizing the empirical cost function

$$\mathbf{D} = \mathcal{F}(\mathbf{Z}) = \arg\min_{\mathbf{D}} f(\mathbf{Z}, \mathbf{D}) = \arg\min_{\mathbf{D}} \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{z}_i, \mathbf{D}) \tag{1}$$

where ℓ is a loss function such that $\ell(\boldsymbol{z}, \mathbf{D})$ approximate to 0 when $\mathbf{D}$ perfectly represent $\boldsymbol{z}$. $\ell(\boldsymbol{z}, \mathbf{D})$ can approximates the sparse solution $\boldsymbol{x}$ by solving the l_1-minimization problem, which yields the convex optimization problem [24, 26]

$$\ell(\boldsymbol{z}, \mathbf{D}) = \min_{\boldsymbol{x} \in \mathbb{R}^k} \frac{1}{2} \|\boldsymbol{z} - \mathbf{D}\boldsymbol{x}\|_2^2 + \lambda \|\boldsymbol{x}\|_1 \tag{2}$$

which can be rewritten as a matrix factorization problem with a sparsity penalty term

$$\ell(\mathbf{Z}, \mathbf{D}) = \min_{\mathbf{D} \in \mathcal{C}, \mathbf{X} \in \mathbb{R}^{k \times n}} \frac{1}{2} \|\mathbf{Z} - \mathbf{D}\mathbf{X}\|_F^2 + \lambda \|\mathbf{X}\|_{1,1} \tag{3}$$

where λ is the regularization parameter and $\mathcal{C}$ is the convex set of matrices verifying this constraints:

$$\mathcal{C} \triangleq \{\mathbf{D} \in \mathbb{R}^{m \times k} \ \text{ s.t. } \forall \ j = 1, \ldots, k, \quad d_j^T d_j \leq 1\} \tag{4}$$

Assuming that the training set is composed of i.i.d. samples of a distribution $p(\boldsymbol{z})$, *i.e.,* $\boldsymbol{z} \sim p(\boldsymbol{z})$, one element $\boldsymbol{z}_t$ is drawn from $\mathbf{Z}$ at a time in the inner loop of the learning process where t is current iteration. Given the dictionary $\mathbf{D}_{t-1}$ obtained from the previous iteration and the sparse code, $\boldsymbol{x}_i \ \forall \ i < t$, computed during the previous iterations, the updated dictionary $\mathbf{D}_t$ are computed by minimizing the following Quadratic Surrogate (QS) function

$$\hat{f}_t(\mathbf{Z}, \mathbf{D}) = \frac{1}{t} \sum_{i=1}^{t} \left(\frac{1}{2} \|\boldsymbol{z}_i - \mathbf{D}\boldsymbol{x}_i\|_2^2 + \lambda \|\boldsymbol{x}_i\|_1 \right) \tag{5}$$

In [26], Mairal *et al.*proved that the QS function $\hat{f}_t$ has approximate upper bound of f in Eq. (1) and can converge to the same limit of f_t. Thus, $\hat{f}_t$ acts as a surrogate for f. As $\hat{f}_t$ is close to $\hat{f}_{t-1}$ for large values of t, so are $\mathbf{D}_t$ and $\mathbf{D}_{t-1}$ under suitable assumptions, which makes it efficient to use $\mathbf{D}_{t-1}$ as warm restart for computing $\mathbf{D}_t$. QS algorithm guarantees that it certainly converges to the set of stationary points of the dictionary learning problem. Without the proof, the convergence of K-SVD is uncertain. Furthermore, QS solver can adapt prior knowledge from past learning processes to improve current dictionary learning, this property is not possessed by K-SVD.

Given the QS function $\hat{f}_t$, Eq. (1) can be rewritten as

$$\mathbf{D} = \mathcal{F}_{\text{qs}}(\mathbf{Z}) = \arg\min_{\mathbf{D}} \hat{f}_t(\mathbf{Z}, \mathbf{D}) \tag{6}$$

The details of dictionary learning and prior knowledge adaptation of $\mathcal{F}_{\text{qs}}$ will be elaborated in Section 4.

Similar to [18], the saliency prediction problem is casted as a binary classification problem in this work. Given the training samples from Section 3.1, a discriminative sparse error term, $\|\mathbf{U} - \mathbf{L}\mathbf{X}\|_F^2$, and a classification error term, $\|\boldsymbol{v}^T - \boldsymbol{w}^T\mathbf{X}\|_2^2$, are taken into account to approximate the discriminative sparse codes $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{k\times n}$ and to learn a sparse dictionary $\mathbf{D}$. The objective function in the dictionary learning problem for visual saliency prediction can be formulated as:

$$\begin{aligned} <\mathbf{D}, \mathbf{L}, \mathbf{X}, \boldsymbol{w}> = \arg\min_{\mathbf{D},\mathbf{A},\mathbf{X},\boldsymbol{w}} &\|\mathbf{Z} - \mathbf{D}\mathbf{X}\|_F^2 + \alpha\|\mathbf{U} - \mathbf{L}\mathbf{X}\|_F^2 \\ &+ \beta\|\boldsymbol{v}^T - \boldsymbol{w}^T\mathbf{X}\|_2^2 + \lambda\|\mathbf{X}\|_{1,1} \end{aligned} \tag{7}$$

and

$$\mathbf{U} = \begin{pmatrix} \mathbf{U}_0^1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathbf{U}_0^S & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{U}_1^1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{U}_1^S \end{pmatrix} \tag{8}$$

where the coefficients α and β control the relative contribution of the discriminative sparse error term and classification error term, respectively. $\boldsymbol{v}$ is saliency labels from the human fixation ground truth and $\boldsymbol{w}$ is the classification weights to reconstruct the ground truth saliency labels. The matrix $\mathbf{U} \in \{0,1\}^{k\times n}$ is the discriminative sparse codes of input $\mathbf{Z}$ and $\mathbf{L} \in \mathbb{R}^{k\times k}$ is a linear transformation matrix to enforce original sparse codes in $\mathbf{X}$ to be more discriminative. Assuming $\mathbf{Z} = (\mathbf{Z}_0^1, \ldots, \mathbf{Z}_0^S, \mathbf{Z}_1^1, \ldots, \mathbf{Z}_1^S)$ is a set of training features where S is the maximal scale and the subscript 0 and 1 indicate that the training features are from non-salient and salient samples, respectively. $\mathbf{U}_0^s,\ s \in \{1, 2, \ldots, S\}$, is generated by the corresponding $\mathbf{Z}_0^s$. For example, if $\mathbf{Z}_0^s$ only contains $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, $\mathbf{U}_0^s$ is a 2×2 all-ones matrix.

To compute the optimal sparse codes $\mathbf{X}$, Eq. (7) can be rewritten as:

$$<\tilde{\mathbf{D}}, \mathbf{X}>= \arg\min_{\tilde{\mathbf{D}},\mathbf{X}} \|\tilde{\mathbf{Z}} - \tilde{\mathbf{D}}\mathbf{X}\|_F^2 + \lambda\|\mathbf{X}\|_{1,1} \tag{9}$$

where $\tilde{\mathbf{Z}}$ and $\tilde{\mathbf{D}}$ are denoted as:

$$\tilde{\mathbf{Z}} = (\mathbf{Z}^T, \sqrt{\alpha}\mathbf{U}^T, \sqrt{\beta}\mathbf{v})^T \tag{10}$$

$$\tilde{\mathbf{D}} = (\mathbf{D}^T, \sqrt{\alpha}\mathbf{L}^T, \sqrt{\beta}\mathbf{w})^T \tag{11}$$

and λ is a regularization parameter. Now Eq. (9) becomes a typical sparse coding problem.

3.3. Saliency Prediction

Given a learned dictionary $\mathbf{D}$ and a set of feature patches $\mathbf{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n] \in \mathbb{R}^{m\times n}$ extracted from all pixels in a test image. The sparse code $\boldsymbol{x}$ and saliency value v for each corresponding $\boldsymbol{z}$ can be computed as follows:

$$\boldsymbol{x} = \arg\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{z} - \mathbf{D}\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_1 \tag{12}$$

$$v = (\boldsymbol{w}^T\boldsymbol{x}) \cdot \left|\boldsymbol{w}^T\boldsymbol{x}\right| \tag{13}$$

where Eq. (12) is solved with LARS algorithm [10] and $\boldsymbol{w}$ is obtained from Eq. (7). The predicted v from each pixel location form a saliency response $\mathbf{M}$.

Finally, to represent the conspicuity at every location in the visual field by a scalar quantity and simulate the field of view of human attention, saliency response $\mathbf{M}$ is convoluted with a Gaussian kernel $\mathbf{g}$ and the normalization saliency map $\tilde{\mathbf{M}}$ is computed as:

$$\tilde{\mathbf{M}} = \frac{\mathbf{M} * \mathbf{g} - \min(\mathbf{M} * \mathbf{g})}{\max(\mathbf{M} * \mathbf{g}) - \min(\mathbf{M} * \mathbf{g})} \tag{14}$$

where $*$ represents the convolution operator.

4. Online Dictionary Learning

In this section, we elaborate the detail steps to solve Eq. (9) with the Label Consistent Quadratic Surrogate (LCQS) algorithm, followed by the online mathematical structure to update the dictionary, and the initialization and optimization of the LCQS model. The proposed online saliency model with the LCQS algorithm is summarized in Algorithm 1.

4.1. Dictionary Learning & Update

Given a set of training samples $\tilde{\mathbf{Z}} = [\tilde{\boldsymbol{z}}_1, \ldots, \tilde{\boldsymbol{z}}_n]$ where $\tilde{\boldsymbol{z}}_i \in p(\tilde{\boldsymbol{z}})$, one sample $\tilde{\boldsymbol{z}}_t$ is drawn from $\tilde{\mathbf{Z}}$, at iteration t, to compute the decomposition of $\tilde{\boldsymbol{z}}_t$, $\boldsymbol{x}_t$, with the dictionary

learned in the previous iteration, $\tilde{\mathbf{D}}_{t-1}$, using LARS algorithm [10]

$$\boldsymbol{x}_t = \arg\min_{\boldsymbol{x}\in\mathbb{R}^k} \frac{1}{2}\|\tilde{\boldsymbol{z}}_{t-1} - \tilde{\mathbf{D}}_{t-1}\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_1 \tag{15}$$

The computed $\boldsymbol{x}_t$ will be used to update the knowledge matrices $\mathbf{Q}$ and $\mathbf{H}$ via

$$\begin{aligned} \mathbf{Q}_t &\leftarrow \mathbf{Q}_{t-1} + \boldsymbol{x}_t\boldsymbol{x}_t^T \\ \mathbf{H}_t &\leftarrow \mathbf{H}_{t-1} + \tilde{\boldsymbol{z}}_t\boldsymbol{x}_t^T \end{aligned} \tag{16}$$

where $\mathbf{Q}_0$ and $\mathbf{H}_0$ are both zero matrices if there is no prior information. At the meantime, the objective function in Eq. (9) can be rewritten in an iterative fashion

$$\begin{aligned} \tilde{\mathbf{D}}_t &= \arg\min_{\tilde{\mathbf{D}}\in\mathcal{C}} \frac{1}{t}\sum_{i=1}^{t}\left(\frac{1}{2}\|\tilde{\boldsymbol{z}}_i - \tilde{\mathbf{D}}\boldsymbol{x}_i\|_2^2 + \lambda\|\boldsymbol{x}_i\|_1\right) \\ &= \arg\min_{\tilde{\mathbf{D}}\in\mathcal{C}} \frac{1}{t}\left(\frac{1}{2}\mathrm{Tr}(\tilde{\mathbf{D}}^T\tilde{\mathbf{D}}\mathbf{Q}_t) - \mathrm{Tr}(\tilde{\mathbf{D}}^T\mathbf{H}_t)\right). \end{aligned} \tag{17}$$

In the dictionary update process, the block-coordinate descent method is applied with $\mathbf{D}_{t-1}$ as warm restarts. The update procedure does not require any parameter to control the learning rate. In addition, it does not store the training samples and sparse codes from the previous iterations, but only the thesaurus matrices $\mathbf{Q}_t = [\mathbf{q}_{1,t}, \ldots, \mathbf{q}_{k,t}]$ and $\mathbf{H}_t = [\mathbf{h}_{1,t}, \ldots, \mathbf{h}_{k,t}]$. In each iteration, each basis in $\tilde{\mathbf{D}}$ is sequentially updated, *i.e.,* updating the j-th basis $\mathbf{d}_j$ at a time while freezing the other ones under the constraint $\mathbf{d}_j^T\mathbf{d}_j \leq 1$. Specifically, $\mathbf{d}_j$ is updated to optimize for Eq. (17)

$$\begin{aligned} \mathbf{y}_j &\leftarrow \frac{1}{\mathbf{Q}_{jj}}(\mathbf{h}_j - \tilde{\mathbf{D}}\mathbf{q}_j) + \mathbf{d}_j \\ \mathbf{d}_j &\leftarrow \frac{1}{max(\|\mathbf{y}_j\|_2, 1)}\mathbf{y}_j \end{aligned} \tag{18}$$

In the LCQS model, as $\boldsymbol{x}_i$ is a sparse vector and the coefficients of $\mathbf{Q}_t$ are often concentrated on the diagonal region, the block-coordinate descent method can be performed efficiently. In the dictionary update process, each basis in $\tilde{\mathbf{D}}$ undergoes the update until a convergence criteria is satisfied [26].

4.2. Initialization

For the LCQS algorithm, $\mathbf{D}_0$, $\mathbf{L}_0$ and $\boldsymbol{w}_0$ are initialized as follows. Given the training samples $\mathbf{Z}$, $\mathbf{D}_0$ can be learned with Eq. (3). For $\mathbf{L}_0$, the multivariate ridge regression model [12] is applied with the quadratic loss and l_2-norm regularization as follows

$$\mathbf{L} = \underset{\mathbf{L}}{\mathrm{argmin}}\|\mathbf{U} - \mathbf{L}\mathbf{X}\|^2 + \lambda_2\|\mathbf{L}\|_2^2 \tag{19}$$

which leads to the following solution

$$\mathbf{L}_0 = (\mathbf{X}\mathbf{X}^T + \lambda_2\mathbf{I})^{-1}\mathbf{X}\mathbf{U}^T \tag{20}$$

Algorithm 1: Pseudo code for Label Consistent Quadratic Surrogate algorithm

Input: $\mathbf{Z}, \mathbf{v}, T, \lambda, \alpha, \beta$
Output: $\mathbf{D}, \boldsymbol{w}$

Initialize: $\boldsymbol{w}_0, \mathbf{L}_0, \mathbf{U}$
if *Prior knowledge exists* **then**
  $\mathbf{Q}_0 \leftarrow \mathbf{Q}_{past}$, $\mathbf{H}_0 \leftarrow \mathbf{H}_{past}$;
else
  $\mathbf{Q}_0 \leftarrow 0$, $\mathbf{H}_0 \leftarrow 0$;
end
Compute $\tilde{\mathbf{Z}}$ and $\tilde{\mathbf{D}}$ by Eq. (10) and Eq. (11);
for $t = 1, 2, \ldots, T$ **do**
  Draw $\tilde{\boldsymbol{z}}_t$ from $\tilde{\mathbf{Z}}$.
  Compute $\boldsymbol{x}_t$ with Eq. (15)
  Compute $\mathbf{Q}_t, \mathbf{H}_t$ with Eq. (16)
  Compute $\tilde{\mathbf{D}}_t$ using block-coordinate descent method with $\tilde{\mathbf{D}}_{t-1}$ as warm restart:
  repeat
    for $j = 1$ *to* k **do**
      Update sequentially the j-th column to optimize by Eq. (18).
    end
  until *convergence*
  Update $\tilde{\mathbf{D}}_t$
end
Decompose $\mathbf{D}$ and $\boldsymbol{w}$ from $\tilde{\mathbf{D}}$ by Eq. (11)
return $\mathbf{D}$ *and* $\boldsymbol{w}$

where $\mathbf{I}$ is an identity matrix and λ_2 is the regularization parameter. Similar to initializing $\mathbf{L}_0$, $\boldsymbol{w}_0$ can be obtained by

$$\boldsymbol{w}_0 = (\mathbf{X}\mathbf{X}^T + \lambda_1\mathbf{I})^{-1}\mathbf{X}\boldsymbol{v}^T \tag{21}$$

where λ_1 is the l_1-norm regularization parameter. Once $\mathbf{D}_0$ is computed, the LARS algorithm is performed to compute $\mathbf{X}$ which will be fed to Eq. (20) and Eq. (21) to initialize $\mathbf{L}_0$ and $\boldsymbol{w}_0$.

4.3. Prior Knowledge Adaptation

As illustrated in Algorithm 1, the thesaurus matrices $\mathbf{Q}$ and $\mathbf{H}$ can be generated and saved as prior knowledge in each iteration. When we initiate a new dictionary training process with an unseen dataset, the proposed model first reviews if there exists prior knowledge generated from the previous dictionary learning processes. If there is no prior knowledge, $\mathbf{Q}$ and $\mathbf{H}$ are initialized as zero matrices. As shown in our experiments, the prior knowledge improves the dictionary learning, especially in the scenario where the training dataset is relatively small.

Table 1: Overview of the eye tracking datasets. The human eye fixation ground truth were collected from free-viewing on each image.

Datasets	Subjects	Durations	Images
MIT [21]	15	3 sec	1003 natural indoor and outdoor scene images
OSIE [37]	15	3 sec	700 natural indoor and outdoor scene images, aesthetic photographs from Flickr and Google
NUSEF [32]	25	5 sec	758 everyday scene images from Flickr, aesthetic content from Photo.net, Google, emotion-evoking IAPS pictures

4.4. Optimization

Leverage the prior knowledge. At each iteration, the new thesaurus information is updated with equal weight as the prior information. In the online learning literature, a general practice is to allocate new information with more weight while reducing the weight of existing information to boost the process of convergence [27]. By taking this practice into account, Eq. (16) can be replaced by

$$\begin{aligned}\mathbf{Q}_t &\leftarrow \beta\mathbf{Q}_{t-1} + \boldsymbol{x}_t\boldsymbol{x}_t^T \\ \mathbf{H}_t &\leftarrow \beta\mathbf{H}_{t-1} + \tilde{\boldsymbol{z}}_t\boldsymbol{x}_t^T\end{aligned} \tag{22}$$

where $\beta = (1 - \frac{1}{t})^{\rho}$ and ρ is the convergence rate factor. Correspondingly, Eq. (17) becomes

$$\begin{aligned}\tilde{\mathbf{D}}_t &= \underset{\tilde{\mathbf{D}}\in\mathcal{C}}{\operatorname{argmin}} \frac{1}{\sum_{j=1}^{t}\left(\frac{j}{t}\right)^{\rho}} \sum_{i=1}^{t}\left(\left(\frac{i}{t}\right)^{\rho}\|\tilde{\boldsymbol{z}}_i - \tilde{\mathbf{D}}\boldsymbol{x}_i\|_2^2 + \lambda\|\boldsymbol{x}_i\|_1\right) \\ &= \underset{\tilde{\mathbf{D}}\in\mathcal{C}}{\operatorname{argmin}} \frac{1}{\sum_{j=1}^{t}\left(\frac{j}{t}\right)^{\rho}}\left(\frac{1}{2}\operatorname{Tr}(\tilde{\mathbf{D}}^T\tilde{\mathbf{D}}\mathbf{Q}_t) - \operatorname{Tr}(\tilde{\mathbf{D}}^T\mathbf{H}_t)\right)\end{aligned} \tag{23}$$

Now, Eq. (17) is a special case of Eq. (23) when $\rho = 0$.

Update with mini-batch. To improve the convergence speed, $\eta > 1$ samples are drawn at each iteration instead of a single sample using the same heuristic in the stochastic gradient descent algorithm. Let us denote $\tilde{\boldsymbol{z}}_{t,1}, \ldots, \tilde{\boldsymbol{z}}_{t,\eta}$ as the samples drawn at iteration t. Hence, Eq. (16) can be rewritten to update the thesaurus information with multiple training samples as:

$$\begin{aligned}\mathbf{Q}_t &\leftarrow \mathbf{Q}_{t-1} + \frac{1}{\eta}\sum_{i=1}^{\eta}\boldsymbol{x}_{t,i}\boldsymbol{x}_{t,i}^T \\ \mathbf{H}_t &\leftarrow \mathbf{H}_{t-1} + \frac{1}{\eta}\sum_{i=1}^{\eta}\tilde{\boldsymbol{z}}_{t,i}\boldsymbol{x}_{t,i}^T\end{aligned} \tag{24}$$

5. Experiments

5.1. Experimental Configuration

Datasets and Evaluation Configuration. The proposed model is evaluated on 3 benchmark eye tracking datasets: MIT dataset (MIT) [21], Object and Semantic Images and Eye-tracking dataset (OSIE) [37] and NUS Eye-Fixation (NUSEF) dataset [32]. These datasets consist of large number of affective stimuli [5], which is beneficial in this work. The details are summarized in Table 1. In this work, we compared the proposed method with the LC-KSVD saliency model [18] and 4 state-of-the-art bottom-up saliency models (*i.e.,* Itti [17], GBVS [13], SUN [39] and Image Signature [14]).

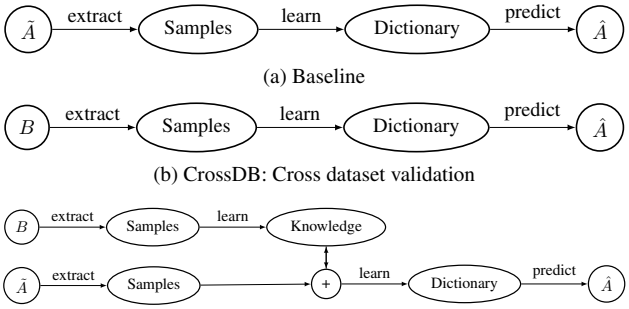

Figure 2: Conceptual illustration of the three experiment configurations employed in this work. A and B are datasets, where A is randomly divided into training set $\tilde{A}$ and evaluation set $\hat{A}$.

Three types of experiment configuration are used to validate the performance. Firstly, we evaluate the proposed model with the conventional learning strategy (denoted as *Baseline*, see Fig. 2(a)). Under this strategy, the training set and evaluation set are both selected from the dataset $\mathbf{A}$. Secondly, we conduct evaluation with cross dataset validation (denoted as *CrossDB*, see Fig. 2(b)). In this configuration, saliency model is trained on dataset $\mathbf{B}$ and predicts on dataset $\mathbf{A}$. Thirdly, we leverage the prior information from another dataset to improved the model's quality. The training and prediction are both conducted on dataset $\mathbf{A}$, while the trained model with the prior knowledge of dataset $\mathbf{B}$ is used under online learning configuration (denoted as *Adaptive*, see Fig. 2(c)).

The correlations of the visual content between datasets are considered for datasets selection for the CrossDB and Adaptive configurations. The MIT and OSIE datasets both contain natural scene images. Hence, we conducted two experiments by mirroring the role of MIT and OSIE, which provides understanding when leveraging prior knowledge from similar dataset as well as study the impact of the center bias factor. We also conduct one set of experiment by leveraging information from MIT to NUSEF in order to exploit the disadvantage of naive CrossDB method. A large

portion of NUSEF images contain emotional faces, nudes and actions, which have semantic impact on human fixations and different from the MIT and OSIE datasets.

Evaluation Metrics. There are several widely used metrics to evaluate the performance of visual saliency models with human fixation data. The Area Under the ROC curve (AUC) [35] considers human fixations as the positive set and some points from the image are randomly chosen as the negative set. However, AUC generates a large value for a central Gaussian model and is affected by center bias [34]. To address this problem, the shuffled AUC (sAUC) [35, 39] was introduced to select negative samples from human fixation locations from all other training sample. In addition, the Normalized Scanpath Saliency (NSS) [31] and the Correlation Coefficient (CC) [30] are employed to measure the performance. NSS is defined as the average saliency value at the fixated locations in the normalized predicted saliency map which has zero mean and unit variance, whereas CC measures the linear correlation between the saliency prediction and the ground truth. As mentioned in [34], observers show a marked tendency to fixate on the screen center and this centre bias is shown in the MIT dataset [21]. The GBVS model implicitly used center-preference to predict saliency [13]. To conduct fair comparison, we use a 200×200 pixels Gaussian blob ($\sigma = 60$) as center bias and multiplying with saliency maps to compute CC and NSS [3].

The Gaussian kernel for blurring affects the sAUC, NSS and CC scores. We parametrize the standard deviation of the blurring kernel from 0 to 0.08 in steps of 0.01. We first generate the saliency maps from various models without smoothing, followed by blurring them with various kernels. The blurred saliency maps are used to generate the respective scores. The three evaluation metrics are complementary and provide a more objective evaluation of various models. All performance is reported with the mean accuracy using 10-fold cross validations.

5.2. Performance Evaluation

The quantitative results are shown in Fig. 3 whereas the maximum scores are shown in Table 2. We first evaluate the results under the baseline configuration. The proposed LCQS-Baseline outperforms all methods across all blur widths with a noticeable margin in most scenarios. This is due to the advantage from the dictionary training stage and the convergence properties of the QS algorithm. On MIT, LCQS-Baseline remarkably outperforms other models on sAUC, whereas the best NSS and CC of the other models are close to LCQS-Baseline. Quite a number of MIT images have a dominant object in the center of the image, this results a saliency model to detect the same object as predicted by other models.

For the cross dataset validation, LCQS-CrossDB significantly outperforms LCKSVD-CrossDB which shows a consistent pattern between LCQS-Baseline and LCKSVD-Baseline. The sAUC, NSS and CC of LCQS-CrossDB is lower than LCQS-Baseline's on OSIE and NUSEF, whereas LCQS-CrossDB is higher than LCQS-Baseline on MIT. This is partly due to the fact that MIT has a considerable portion of fixations in the center and the saliency map of LCQS-CrossDB has more false detections on the center.

Table 2: Qualitative results of the proposed LCQS saliency model and various state-of-the-art models. The accuracy is measured with the shuffled AUC and reported results is the mean accuracy with 10-fold validations. The best performance on each dataset are in **BOLD**.

	A=MIT, B=OSIE	A=OSIE, B=MIT	A=NUSEF, B=MIT
Itti	0.6271	0.6575	0.5816
SUN	0.6609	0.7353	0.6172
Signature	0.6795	0.7487	0.6267
GBVS	0.6694	0.7055	0.6112
LCKSVD - Baseline	0.6846	0.7479	0.6406
LCKSVD - CrossDB	0.6694	0.7127	0.6374
LCQS - Baseline	0.6898	0.7649	0.6495
LCQS - CrossDB	0.6935	0.7415	0.6379
LCQS - Adaptive	**0.7012**	**0.7696**	**0.6517**

LCQS-Adaptive achieved the best performance over sAUC, NSS and CC across all blur widths on all datasets. Compared to the naive cross dataset configuration, it benefits from eliminating the dataset bias by leveraging other datasets' prior knowledge. The improvement of LCQS-Adaptive over LCQS-Baseline on NUSEF is also observed (with smaller margin) despite that the visual content on MIT are significant different.

The qualitative results are shown in Fig. 4. LCQS-Baseline shows more consistent maps with human fixations than other comparative models. For example, it better detects the two ships in the fourth row. By taking the prior knowledge, LCQS-Adaptive has a stronger response on human face in the sixth row which better approximates human fixations than the result of LCQS-Baseline.

6. Conclusions

In this work, we have presented a new learning based saliency prediction model, which employ an iterative online algorithm to learn a sparse dictionary with label consistent constraints. By utilizing the advantage of quadratic surrogate algorithm and label consistent constrains, the proposed model consistently achieves noticeable improvement over existing state-of-the-art saliency models, as well as addressing the problem of insufficient eye fixation datasets by leveraging the prior knowledge from a learned model to improve the quality of learning.

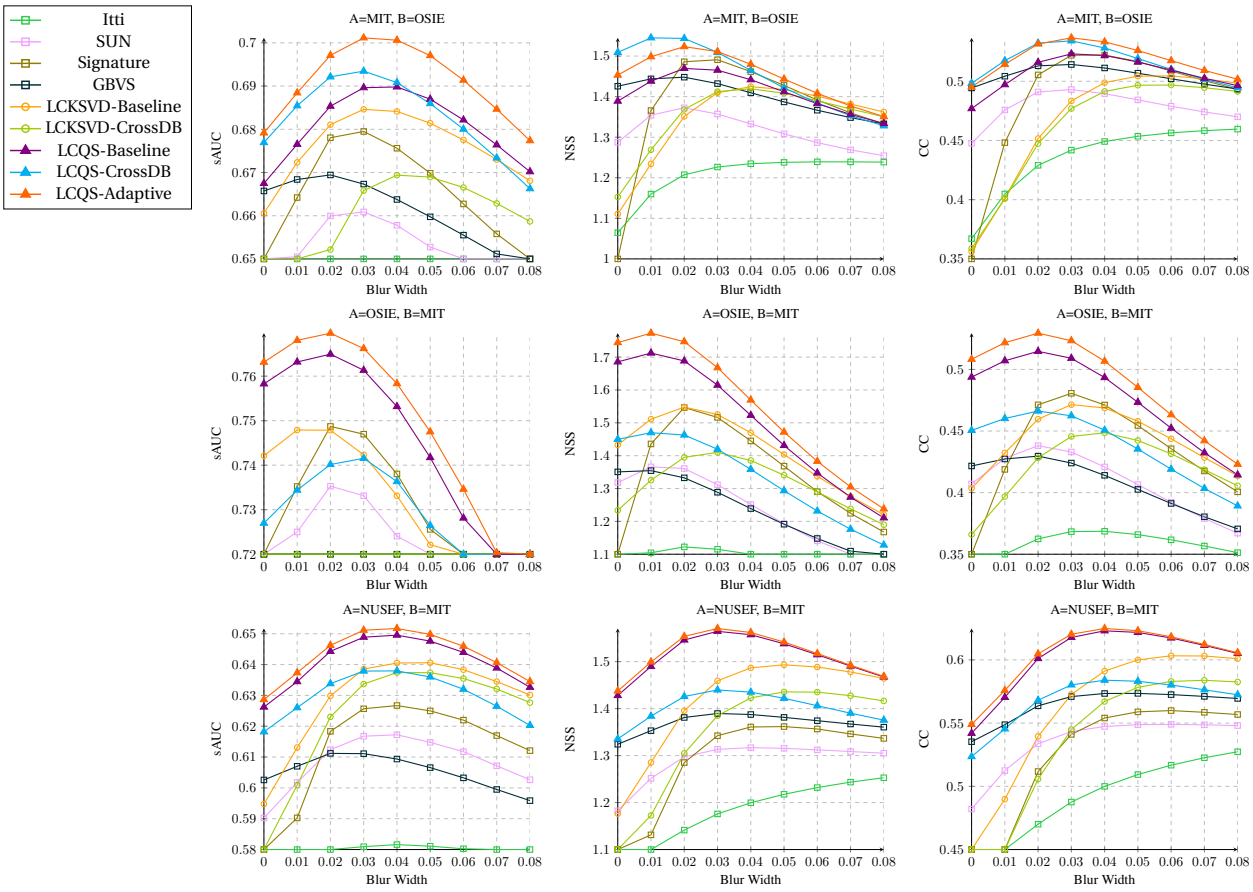

Figure 3: Shuffled AUC, NSS and CC scores on various datasets with various models. The blur width is parameterized by a Gaussian's standard deviation in image widths.

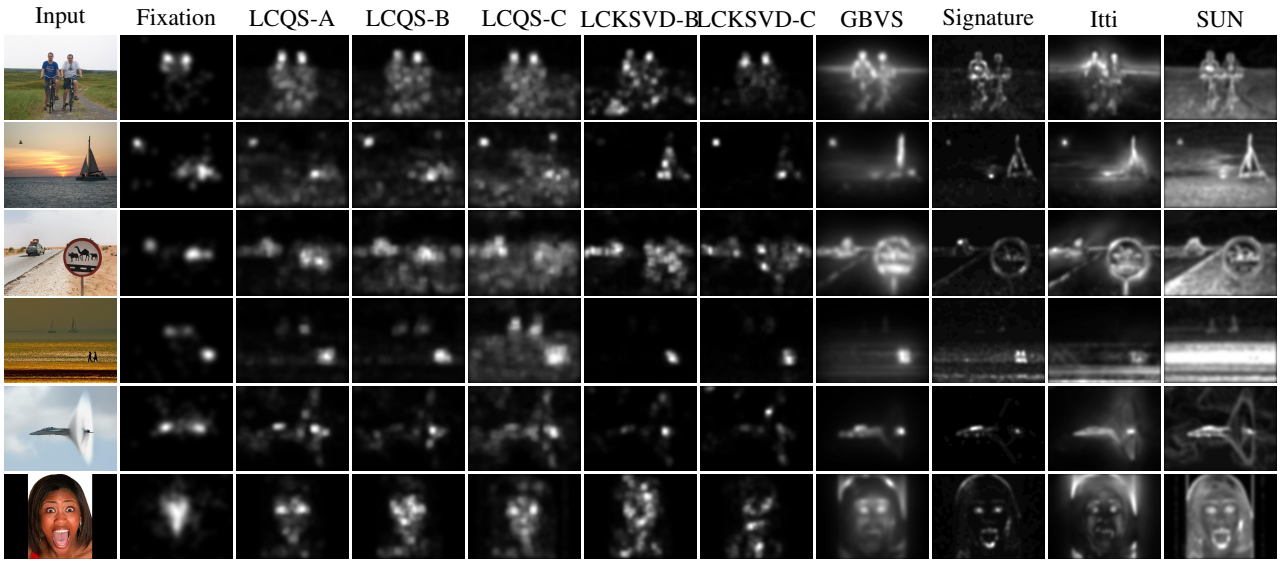

Figure 4: Qualitative results of the proposed LCQS saliency model and various state-of-the-art models. Row 1 & 2 are samples from MIT dataset, Row 3 & 4 are samples from OSIE dataset, and Row 5 & 6 are samples from NUSEF dataset. The configuration for the cross dataset learning are as stated in Section 5.1. The suffix *A*, *B*, and *C* stand for Adaptive, Baseline, and CrossDB, respectively.

Acknowledgments

The research was supported by the Singapore NRF under its IRC@SG Funding Initiative and administered by the IDMPO, the Defense Innovative Research Programme (No. 9014100596), and the Ministry of Education Academic Research Fund Tier 1 (No. R-263-000-A49-112).

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.

[2] B. Babenko, P. Dollár, and S. J. Belongie. Task specific local region matching. In *IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007*, pages 1–8, 2007.

[3] A. Borji. Boosting bottom-up and top-down visual features for saliency estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 438–445, 2012.

[4] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207, 2013.

[5] A. Borji, H. Tavakoli, D. Sihite, and L. Itti. Analysis of scores, datasets, and models in visual saliency prediction. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 921–928, Dec 2013.

[6] K. Crammer, M. Dredze, and F. Pereira. Exact convex confidence-weighted learning. In *Advances in Neural Information Processing Systems*, pages 345–352, 2008.

[7] K. Crammer and D. D. Lee. Learning via Gaussian herding. In *Advances in Neural Information Processing Systems*, pages 451–459, 2010.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005.

[9] J. C. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

[10] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[11] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[12] G. H. Golub, P. C. Hansen, and D. P. O'Leary. Tikhonov regularization and total least squares. *SIAM Journal on Matrix Analysis and Applications*, 21(1):185–194, 1999.

[13] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems*, pages 545–552, 2006.

[14] X. Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):194–201, 2012.

[15] L. Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing*, 13(10):1304–1318, 2004.

[16] L. Itti and C. Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001.

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

[18] M. Jiang, M. Song, and Q. Zhao. Leveraging human fixations in sparse coding: Learning a discriminative dictionary for saliency prediction. In *IEEE International Conference on Systems, Man., and Cybernetics*, pages 2126–2133, 2013.

[19] Z. Jiang, Z. Lin, and L. S. Davis. Learning a discriminative dictionary for sparse coding via label consistent K-SVD. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1697–1704. IEEE, 2011.

[20] Z. Jiang, Z. Lin, and L. S. Davis. Label consistent K-SVD: learning a discriminative dictionary for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2651–2664, 2013.

[21] T. Judd, K. A. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *IEEE International Conference on Computer Vision*, pages 2106–2113, 2009.

[22] J. Kivinen, A. J. Smola, and R. C. Williamson. Online learning with kernels. *IEEE Transactions on Signal Processing*, 52(8):2165–2176, 2004.

[23] C. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. *Matters of Intelligence*, 188:115–141, 1987.

[24] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems*, pages 801–808, 2007.

[25] V. Mahadevan and N. Vasconcelos. Saliency-based discriminant tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1007–1013, 2009.

[26] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:19–60, 2010.

[27] R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. *Learning in Graphical Models*, pages 355–368, 1998.

[28] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson. Top-down control of visual attention in object detection. In *ICIP (1)*, pages 253–256, 2003.

[29] F. Orabona and K. Crammer. New adaptive algorithms for online classification. In *Advances in Neural Information Processing Systems*, pages 1840–1848, 2010.

[30] 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*, 3(1):13–24, 2004.

[31] R. J. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(18):2397–2416, 2005.

[32] S. Ramanathan, H. Katti, N. Sebe, M. S. Kankanhalli, and T.-S. Chua. An eye fixation database for saliency detection in images. In *Lecture Notes in Computer Science*, volume 6314, pages 30–43, 2010.

[33] U. Rutishauser, D. Walther, C. Koch, and P. Perona. Is bottom-up attention useful for object recognition? In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 37–44, 2004.

[34] 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*, 7(14), 2007.

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

[36] J. Wang, P. Zhao, and S. C. Hoi. Exact soft confidence-weighted learning. In *International Conference on Machine Learning*, pages 121–128, 2012.

[37] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao. Predicting human gaze beyond pixels. *Journal of Vision*, 14(1):1–20, 2014.

[38] J. Zhang and S. Sclaroff. Saliency detection: A Boolean map approach. In *IEEE International Conference on Computer Vision*, pages 153–160, 2013.

[39] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):32, 2008.

[40] Q. Zhao and C. Koch. Learning a saliency map using fixated locations in natural scenes. *Journal of vision*, 11(3):9, 2011.