Top-Down Visual Saliency via Joint CRF and Dictionary Learning

Jimei Yang and Ming-Hsuan Yang

Abstract—Top-down visual saliency is an important module of visual attention. In this work, we propose a novel top-down saliency model that jointly learns a Conditional Random Field (CRF) and a visual dictionary. The proposed model incorporates a layered structure from top to bottom: CRF, sparse coding and image patches. With sparse coding as an intermediate layer, CRF is learned in a feature-adaptive manner; meanwhile with CRF as the output layer, the dictionary is learned under structured supervision. For efficient and effective joint learning, we develop a max-margin approach via a stochastic gradient descent algorithm. Experimental results on the Graz-02 and PASCAL VOC datasets show that our model performs favorably against state-of-the-art top-down saliency methods for target object localization. In addition, the dictionary update significantly improves the performance of our model. We demonstrate the merits of the proposed top-down saliency model by applying it to prioritizing object proposals for detection and predicting human fixations.

Index Terms—Visual saliency, top-down visual saliency, fixation prediction, dictionary learning and conditional random fields

1 Introduction

Visual saliency has attracted much attention in the vision community and numerous computational models have been proposed. Early work focuses on its bottom-up process and establishes a number of saliency principles such as center-surround contrast [1], self-information [2], topological connectivity [3] and spectral residual [4]. Central to these principles are the measures of abnormality or distinctiveness of one image region within a context. As a result, bottom-up saliency maps are shown to be effective in simple scenes for predicting human fixations [1], [2], [3], [5] and for highlighting the informative regions of images [4], [6].

In this paper, we investigate top-down visual saliency, complementary to bottom-up visual saliency for visual attention [7], [8]. Top-down models, similar to bottom-up ones, are also based on local image evidences within their contexts. However, different from bottom-up models, top-down models are driven not only by image contexts but also by specific visual priors. We define top-down visual saliency as the distinctiveness of target objects from their surroundings within an image. The goal of top-down saliency detection is to highlight the target objects and suppress the backgrounds.

The advantages of top-down models become more clear when they are applied to complex scenes, where bottom-up saliency models usually respond to numerous unrelated low-level visual stimuli (i.e., false positives) and miss the objects of interest (i.e., false negatives) due to the nature of data-driven formulations.

We propose a novel top-down visual saliency model based on image patches. The goal of our model is to learn from labeled training examples from a number of classes to localize target objects in an image. We use a binary variable to indicate the presence or absence of a target object in an image patch. The saliency value of an image patch is computed by the probability of a target object being present at that location. We formulate the saliency model with a layered conditional random field (CRF) model in which target variables are conditioned on sparse codes of image patches. The use of a conditional random field enables us to exploit the connectivity of adjacent image patches such that the saliency map is computed by incorporating local context information. Meanwhile, the use of sparse coding facilitates us to model feature selectivity for target prediction, which typically results in a more compact and discriminative representation. The presence of target objects in an image can be thus inferred by message passing, and represented by posterior probabilities. We compute the saliency map by normalizing those posterior probabilities of patches within their context, thereby turning it to be a context-dependent image attribute.

We note that the proposed model is more than a straightforward combination of CRF and sparse coding. Instead, it accommodates joint CRF and dictionary learning. By using sparse codes as intermediate layer, we learn a both a discriminative dictionary under the supervision of CRF, and a CRF model driven by sparse coding. We propose a max-margin approach to train the model by exploiting fast inference algorithms such as the graph cut method [9].

We apply the learned top-down saliency maps to prioritizing object proposals and for predicting human fixations. The state-of-the-art object detection and segmentation

- *J. Yang is with Adobe Research, San Jose, CA. E-mail: jimyang@adobe.com.*
- *M.-H. Yang is with School of Engineering, University of California, Merced, CA. E-mail: mhyang@ucmerced.edu.*

Manuscript received 3 Jan. 2015; revised 31 Jan. 2016; accepted 22 Feb. 2016. Date of publication 27 Mar. 2016; date of current version 13 Feb. 2017. Recommended for acceptance by C. Sminchisescu.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2016.2547384

algorithms [10], [11] rely on evaluating highly complex features on numerous candidate image regions (e.g., object proposals [12], [13]). Given the heavy computational load, it is desirable to prioritize highly plausible object proposals over the others such that early decisions can be made for accelerating detection or segmentation tasks. The category-specific top-down saliency maps can naturally be used to rank all the object proposals. (i.e., early stages in the object detection and segmentation processes). In addition, when integrating category-specific top-down saliency maps with image classifiers in a probabilistic sense, we obtain category-independent top-down saliency maps as generic *objectness* measures [14] to highlight image regions of interest. We combine them with complimentary bottom-up saliency maps for predicting human fixations.

We present results on the Graz-02 [15] and the PASCAL VOC [16] datasets. On the Graz-02 dataset, our model demonstrates promising performance against two state-of-the-art top-down saliency algorithms [8], [17] for object localization. On the PASCAL VOC dataset, we train top-down saliency models for localizing the target objects of 20 different classes, and present cross-category saliency analysis that reveals the affinity relationship among different object categories in terms of both local appearance similarity and co-occurrence. We use our category-specific top-down saliency maps to rank object proposals generated by selective search [13], and obtain above 90 percent average recall rates at 1,000 proposals on 20 object classes. We calculate category-independent top-down saliency maps by integrating 20 class-specific maps with state-of-the-art object classifiers [18]. We present fixation prediction experiments on the PASCAL_S dataset (the validation set of PASCAL VOC 2010) by combining our top-down saliency maps with bottom-up saliency maps [3], [5]. The results show that our method performs favorably against the state-of-the-art fixation prediction algorithms.

2 Related Work

We discuss the related algorithms for top-down saliency maps and human fixation prediction. In addition, we briefly overview the relevant CRF and dictionary learning methods.

2.1 Top-Down Saliency Maps

Gao et al. [17] propose a top-down saliency algorithm by selecting discriminant features from a pre-defined filter bank. The discriminant features are characterized by the statistical difference based on the presence or absence of features in the object class of interest. With the selected features, the saliency values of interest points can be computed based on their mutual information. Instead of using a pre-defined filter bank, Kanan et al. [8] propose to learn filter responses with independent component analysis (ICA) from natural images. They thus build the top-down component of their saliency model by training a support vector machine (SVM) on ICA features. In our model, the discriminant features are selected from a learned dictionary by sparse coding. In [19], the top-down saliency map is formulated as contextual guidance for object search. This contextual prior performs well when there is a strong correlation between the target locations and holistic scenes, such as cars in urban scenes. However, as target objects are likely to appear anywhere in a scene, this contextual prior is less effective (e.g., images from the Graz-02 and PASCAL VOC datasets). In contrast, we compute the saliency map by inference on a CRF model, which is more effective for incorporating the local context information. Mathe and Sminchisescu [20] propose a dynamic top-down saliency algorithm to predict human eye movements when looking at actions and contexts.

2.2 Fixation Prediction

Predicting human fixations usually involve both bottom-up saliency maps and top-down modules. Recent methods can be categorized into two approaches. Bayesian visual attention models consider joint probability of objects, features and locations [7], [8]. By applying the Bayes' rule and assuming independence of features and locations, fixation prediction of salient regions is decomposed into feature-driven bottom-up saliency, appearance based top-down saliency and location prior. In [7], a joint probabilistic model is proposed where both bottom-up and top-down saliency can be derived and inference is carried out by message passing. Compared to the Bayesian approach, our method also combines bottom-up and top-down saliency maps, but our model allows hierarchical prediction from part-level features (sparse coding), particular category (class-specific saliency maps) to generic objects (class-independent saliency maps). On the other hand, learning based approaches [21], [22] directly train discriminative classifiers using various features as input and fixation locations as output. Judd et al. [21] present a SVM based method by using different bottom-up image cues and pre-trained object detectors (face and human) as features. In [22], Xu et al. use high-level features for fixation prediction where object-level features from ground truth object masks and semantic-level features from attribute annotations are exploited.

2.3 Conditional Random Fields

CRF models have been successfully applied to various structured output prediction problems such as object segmentation [23] and semantic segmentation [24] due to their flexibility in combining object appearance with context. Previous algorithms [23], [24] usually incorporate CRFs with pre-trained part-based object detectors or bag-of-words classifiers. In this work, we use CRF to generate precise and smooth saliency maps by taking both local appearance and image context into account. Different from [23], [24] that use pre-trained appearance models, our method jointly optimizes the CRF weights and features in local appearance. From this perspective, the proposed algorithm can be extended to semantic segmentation by constructing graphs on superpixels or regions. Note that our model is different from the hidden CRFs [25], where Quattoni et al. define a CRF of latent variables to represent the part features of local patches and a single output variable to describe the image category. Inference is carried out by measuring the compatibility between the image label and the latent variables. In contrast, our model uses sparse coding as latent variables to represent local observations and uses CRF as structured output variables to define the top-down saliency map. Recently, Jain et al. [26] model the joint probability of labels and latent variables with a single CRF energy function for object categorization and segmentation. Tao et al. [27]

further extends this method for semantic segmentation. Our model is based on a layered structure and thus admits efficient back-propagation learning and feed-forward inference without complex joint inference of labeling and visual word assignments.

2.4 Dictionary Learning and Sparse Coding

Recent advances in machine learning facilitate training task-specific dictionaries in a supervised manner [28], [29], [30]. Mairal et al. [28] combine sparse coding and logistic regression into a single loss function. Although this method shows promising results on several vision tasks, it is not clear how it can be effectively applied to complex object recognition problems as the objective function does not take image structures into account. Yang et al. [29] propose a supervised sparse coding method with a hierarchical model for image classification. This method performs well for face recognition as a translation invariant sparse representation is learned with max pooling. In contrast, our model learns discriminative dictionaries with structured output in random fields, It can better capture local context of images for consistent saliency prediction. A recent work [31] also investigates label consistency for supervised dictionary learning. Different from our work, it enforces assignment consistency of visual words during sparse coding. Recently, deep convolutional networks [18] have demonstrated superior performance than sparse coding for feature learning. Learning deep convolutional networks with structured output become increasingly important for dealing with complex visual tasks. In a broad view, our work can be considered as an early attempt for feature learning with structured output.

3 Top-Down Visual Saliency Model

Top-Down visual saliency usually involves object localization [22] and rapid scene understanding [19] from images. Our top-down visual saliency algorithm consists of three stages:

1) *feature extraction:* sparse coding from image patches;
2) *target prediction:* predict target presence with a conditional random field;
3) *activation normalization:* normalize the prediction probabilities in proper context.

The core of this algorithm is the first two steps as the normalization step is application dependent. We unify feature extraction and target prediction into a novel layered model that enjoys joint training of sparse coding and conditional random field. We first introduce the proposed layered model and its joint training, and then describe its application to fixation prediction.

3.1 Patch Based Image Representation

Given any image such as that shown in Fig. 1, we would like to know where the objects of interest lie. We sample a dense grid of patches $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ from the image as the observations. For a local image patch $\mathbf{x} \in \mathbb{R}^p$, we assign a binary label $\mathbf{y}$ to indicate the presence ($\mathbf{y} = 1$) or absence ($\mathbf{y} = -1$) of a target object. The corresponding labels $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots \mathbf{y}_m\}$ carry the information of target presence. The causal relationship between $\mathbf{X}$ and $\mathbf{Y}$ is modeled by the probability $p(\mathbf{Y}|\mathbf{X})$. However, directly inferring the presence of the target from $\mathbf{x}_j$ usually contains only partial information about the target object, resulting in semantic and geometric ambiguities in patch-based representation. It is thus challenging to directly infer the presence of the target from $\mathbf{x}_i$ without considering the others due to the semantic and geometric ambiguities of patch-based representations.

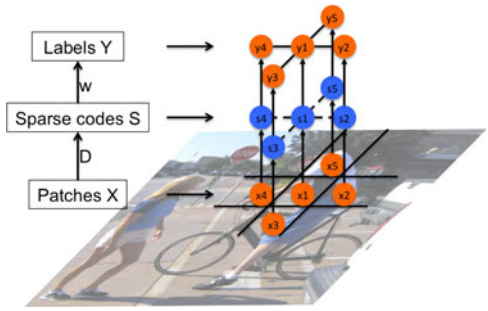


Fig. 1. Proposed model. We construct a layered model on image patches for top-down visual saliency. In the bottom layer, we represent image patches $\mathbf{X}$ with the sparse codes $\mathbf{S}$, using the dictionary $\mathbf{D}$. In the top layer, the binary variables $\mathbf{Y}$, which predict the target presence, form a Markov random field conditioned on sparse codes $\mathbf{S}$, where the pairwise potentials impose the smoothness of label prediction. To learn the model, we develop a max-margin approach to deal with the partition function in the top layer such that the CRF parameters $\mathbf{w}$ and the dictionary $\mathbf{D}$ are learned jointly.

3.2 A Layered Prediction Model

Suppose that there exists a dictionary $\mathbf{D} \in \mathbb{R}^{p \times K}$ that stores the most representative parts (visual words) $\{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_K\}$ learned from the training data. We introduce a vector of latent variables $\mathbf{s}_i \in \mathbb{R}^K$ to compactly represent $\mathbf{x}$ with the dictionary $\mathbf{D}$ by $\mathbf{x}_i = \mathbf{D}\mathbf{s}_i$ by solving the following problem:

$$\mathbf{s}(\mathbf{x}, \mathbf{D}) = \arg\min_{\mathbf{s}} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\mathbf{s}\|^2 + \lambda\|\mathbf{s}\|_1, \tag{1}$$

where λ is a regularization constant. We denote the latent variables for all the patches by $\mathbf{S}(\mathbf{X}, \mathbf{D}) = [\mathbf{s}(\mathbf{x}_1, \mathbf{D}), \mathbf{s}(\mathbf{x}_2, \mathbf{D}), \ldots, \mathbf{s}(\mathbf{x}_m, \mathbf{D})]$. Note that we use the notations $\mathbf{s}(\mathbf{x}, \mathbf{D})$ and $\mathbf{S}(\mathbf{x}, \mathbf{D})$ to emphasize that the sparse latent variables are a function of the dictionary. In the following sections, we introduce $\mathbf{s}_i \triangleq \mathbf{s}(\mathbf{x}, \mathbf{D})$ and $\mathbf{S} \triangleq \mathbf{S}(\mathbf{x}, \mathbf{D})$ to simplify the notations. The sparse coding problem in (1) can be solved efficiently for a single patch by the feature-sign algorithm in [32]. Since the dictionary bases represent the object parts, the sparse code $\mathbf{s}$ contains mid-level representation, i.e., part of a target object in a patch.

If a local patch shows evidence of an object part, it is likely that nearby patches also exhibit similar support. We construct a four-connected graph $\mathcal{G} = <\mathcal{V}, \mathcal{E}>$ on the sampled patches, where $\mathcal{V}$ denote the nodes and $\mathcal{E}$ the edges. Assuming that the labels $\mathbf{Y}$ enjoy the Markov property on the graph $\mathcal{G}$ conditioned on the sparse latent variables $\mathbf{S}(\mathbf{X}, \mathbf{D})$, we formulate a novel CRF model by

$$p(\mathbf{Y}|\mathbf{S}(\mathbf{X}, \mathbf{D}), \mathbf{w}) = \frac{1}{Z} e^{-E(\mathbf{S}(\mathbf{X}, \mathbf{D}), \mathbf{Y}, \mathbf{w})}, \tag{2}$$

where Z is the partition function, $E(\mathbf{S}(\mathbf{X}, \mathbf{D}), \mathbf{Y}, \mathbf{w})$ is the energy function and $\mathbf{w}$ is the CRF weight. This formulation enables us to jointly learn CRF weight $\mathbf{w}$ and the dictionary $\mathbf{D}$. A graphical diagram is shown in Fig. 1. Given the CRF

weight $\mathbf{w}$, the model in (2) can be viewed as CRF modulated dictionary learning, whereas given the dictionary $\mathbf{D}$, it can be viewed as CRF learning with sparse coding. In this model, we predict the presence of targets at a particular node $i \in \mathcal{V}$ from its marginal probability by message passing through the graph

$$p(\mathbf{y}_i|\mathbf{s}_i, \mathbf{w}) = \sum_{\mathbf{y}_{\mathcal{N}(i)}} p(\mathbf{y}_i, \mathbf{y}_{\mathcal{N}(i)}|\mathbf{s}_i, \mathbf{w}), \tag{3}$$

where $\mathcal{N}(i)$ denotes the neighbors of node i on the graph $\mathcal{G}$.

In this work, we decompose the energy function $E(\mathbf{S}(\mathbf{X}, \mathbf{D}), \mathbf{Y}, \mathbf{w})$ into node and pairwise energy terms. For each node $i \in \mathcal{V}$, the energy is measured by the total contribution of sparse codes $\psi(\mathbf{s}_i, \mathbf{y}_i, \mathbf{w}_1) = -\mathbf{y}_i \mathbf{w}_1^\top \mathbf{s}_i$, where $\mathbf{w}_1 \in \mathbb{R}^k$ is the weight vector. For each edge $(i, j) \in \mathcal{E}$, we only consider data-independent smoothness $\psi(\mathbf{y}_i, \mathbf{y}_j, \mathbf{w}_2) = \mathbf{w}_2 \mathbb{I}(\mathbf{y}_i, \mathbf{y}_j)$, where the scalar $\mathbf{w}_2$ measures the weight of labeling smoothness and $\mathbb{I}(\mathbf{y}_i, \mathbf{y}_j)$ is an indicator equaling to one when $\mathbf{y}_i$ and $\mathbf{y}_j$ are different. Therefore, the random field energy can be formulated by

$$E(\mathbf{S}, \mathbf{Y}, \mathbf{w}, \mathbf{D}) = \sum_{i \in \mathcal{V}} \psi(\mathbf{s}_i, \mathbf{y}_i, \mathbf{w}_1) + \sum_{(i,j) \in \mathcal{E}} \psi(\mathbf{y}_i, \mathbf{y}_j, \mathbf{w}_2). \tag{4}$$

Note that our energy function is linear with the parameter $\mathbf{w} = [\mathbf{w}_1; \mathbf{w}_2]$ which is similar to most CRF models [23], [24], [33], but is nonlinear with the dictionary $\mathbf{D}$ that is implicitly defined by $\mathbf{s}(\mathbf{x}, \mathbf{D})$ in (1). This nonlinear parametrization makes it challenging to train the proposed model. We discuss our learning approach in the next section.

Once the optimal CRF parameters $\hat{\mathbf{w}}$ and the dictionary $\hat{\mathbf{D}}$ are learned, a saliency map can be computed efficiently. Our top-down saliency formulation in (2) does not involve complex evaluations of latent variables [25], [26], [28], which makes it feasible to infer the saliency map in a feed-forward manner without alternating between the evaluation of latent variables and label inference.

3.3 Activation Normalization

We define the saliency value of a patch i as the normalized probability of target presence in an image,

$$\mathbf{o}_i = \frac{p(\mathbf{y}_i = 1|\mathbf{s}_i, \mathbf{w})}{\max_{j \in \mathcal{V}} p(\mathbf{y}_j = 1|\mathbf{s}_j, \mathbf{w})}, \tag{5}$$

and accordingly the saliency map is given by $\mathbf{O}(\mathbf{S}, \mathbf{w}) = \{\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_m\}$. This probabilistic definition of a top-down saliency map leverages not only appearance information [8], [17], but also local contextual information through the marginalization in (3).

For a test image $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$, we compute its saliency map $\mathbf{O}$ as follows:

1) solve sparse coding $\mathbf{S}(\mathbf{X}, \hat{\mathbf{D}})$ by (1);
2) compute the posterior probability $p(\mathbf{Y} = 1|\mathbf{S}, \hat{\mathbf{w}})$ by (3);
3) compute the saliency map $\mathbf{O}$ by normalizing the probability (5);
4) upsample $\mathbf{O}$ to the size of test image; optionally blur it with a Gaussian kernel.

4 Joint CRF and Dictionary Learning

Let $\mathcal{X} = \{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots \mathbf{X}^{(N)}\}$ be a set of training instances and $\mathcal{Y} = \{\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \ldots \mathbf{Y}^{(N)}\}$ be the corresponding labels. We aim to learn the CRF parameters $\hat{\mathbf{w}}$ and the dictionary $\hat{\mathbf{D}}$ to maximize the joint likelihood of training samples,

$$\max_{\mathbf{w} \in \mathbb{R}^{(k+1)}, \mathbf{D} \in \mathcal{D}} \prod_{n=1}^{N} p(\mathbf{Y}^{(n)}|\mathbf{S}(\mathbf{X}^{(n)}, \mathbf{D}), \mathbf{w}), \tag{6}$$

where $\mathbf{S}^{(n)}$ is a shorthand of $\mathbf{S}(\mathbf{X}^{(n)}, \mathbf{D})$ and $\mathcal{D}$ is the convex set of dictionaries that satisfies the following constraint:

$$\mathcal{D} = \{\mathbf{D} \in \mathbb{R}^{p \times k}, \|\mathbf{d}_j\|_2 \leq 1, \forall j = 1, 2, \ldots, k\}. \tag{7}$$

4.1 Max-Margin Approach

The difficulties in CRF learning mainly come from evaluating the partition function Z of (2). Motivated by the max-margin CRF learning approaches [23], [33], we pursue the optimal $\mathbf{w}$ and $\mathbf{D}$ such that for all $\mathbf{Y} \neq \mathbf{Y}^{(n)}, n = 1, \ldots, N$

$$p(\mathbf{Y}^{(n)}|\mathbf{S}(\mathbf{X}^{(n)}, \mathbf{D}), \mathbf{w}) \geq p(\mathbf{Y}|\mathbf{S}(\mathbf{X}^{(n)}, \mathbf{D}), \mathbf{w}). \tag{8}$$

This constrained optimization allows us to cancel the partition function Z from both sides of the constraints and express them in terms of energies

$$E(\mathbf{Y}^{(n)}, \mathbf{S}^{(n)}, \mathbf{w}) \leq E(\mathbf{Y}, \mathbf{S}^{(n)}, \mathbf{w}). \tag{9}$$

Furthermore, we expect the ground truth energy $E(\mathbf{Y}^{(n)}, \mathbf{S}(\mathbf{X}^{(n)}, \mathbf{D}), \mathbf{w})$ is less than any other energies $E(\mathbf{Y}, \mathbf{S}(\mathbf{X}^{(n)}, \mathbf{D}), \mathbf{w})$ by a large margin $\Delta(\mathbf{Y}, \mathbf{Y}^{(n)})$. We thus have a new constraint set

$$E(\mathbf{Y}^{(n)}, \mathbf{S}^{(n)}, \mathbf{w}) \leq E(\mathbf{Y}, \mathbf{S}^{(n)}, \mathbf{w}) - \Delta(\mathbf{Y}, \mathbf{Y}^{(n)}). \tag{10}$$

In this paper, we define the margin function $\Delta(\mathbf{Y}, \mathbf{Y}^{(n)}) = \sum_{i=1}^{m} \mathbb{I}(\mathbf{y}_i, \mathbf{y}_i^{(n)})$. There is an exponentially large number of constraints with respect to labeling $\mathbf{Y}^{(n)}$ for each training sample. Similar to the cutting plane algorithm [34], we seek for the most violated constraints by solving

$$\hat{\mathbf{Y}}^{(n)} = \arg\min_{\mathbf{Y}} E(\mathbf{Y}, \mathbf{S}^{(n)}, \mathbf{w}) - \Delta(\mathbf{Y}, \mathbf{Y}^{(n)}). \tag{11}$$

Therefore, we are able to learn the weight $\mathbf{w}$ and the dictionary $\mathbf{D}$ by minimizing the following objective function:

$$\min_{\mathbf{w}, \mathbf{D} \in \mathcal{D}} \frac{\gamma}{2} \|\mathbf{w}\|^2 + \sum_{n=1}^{N} \ell^n(\mathbf{w}, \mathbf{D}), \tag{12}$$

where $\ell^n(\mathbf{w}, \mathbf{D}) \triangleq E(\hat{\mathbf{Y}}^{(n)}, \mathbf{S}^{(n)}, \mathbf{w}) - E(\mathbf{Y}^{(n)}, \mathbf{S}^{(n)}, \mathbf{w})$ and γ controls the regularization of $\mathbf{w}$.

We note that our approach shares a similar objective function with the latent structural SVM [35]. The difference is that the latent structural SVM is linearly parameterized while ours is nonlinear with the dictionary $\mathbf{D}$.

4.2 Learning Algorithm

We propose a stochastic gradient descent algorithm for optimizing the objective function in (12). The basic idea is simple and easy to implement. At the tth iteration, we randomly select a training instance $(\mathbf{X}^{(n)}, \mathbf{Y}^{(n)})$, and then

1) evaluate the sparse latent variables with the dictionary $\mathbf{D}^{(t-1)}$ by (1),
2) obtain the most violated labeling with the weight $\mathbf{w}^{(t-1)}$ by (11),
3) update the weight $\mathbf{w}^{(t)}$ and the dictionary $\mathbf{D}^{(t)}$ by the gradients of the loss function ℓ^n.

We next describe the methods of computing the gradients with respect to the weight and the dictionary.

When the latent variables $\mathbf{S}$ are known, the energy function $E(\mathbf{Y}, \mathbf{S}, \mathbf{w})$ is linear with $\mathbf{w}$ (see (4)),

$$E(\mathbf{Y}, \mathbf{S}, \mathbf{w}) = <\mathbf{w}, f(\mathbf{S}, \mathbf{Y})>, \tag{13}$$

where $f(\mathbf{S}, \mathbf{Y}) = [-\sum_{i\in\mathcal{V}} \mathbf{s}_i\mathbf{y}_i; \sum_{(i,j)\in\mathcal{E}} \mathbb{I}(\mathbf{y}_i, \mathbf{y}_j)]$. We can thus compute the gradient with respect to $\mathbf{w}$,

$$\frac{\partial \ell^n}{\partial \mathbf{w}} = f(\mathbf{S}^{(n)}, \hat{\mathbf{Y}}^{(n)}) - f(\mathbf{S}^{(n)}, \mathbf{Y}^{(n)}) + \gamma\mathbf{w}. \tag{14}$$

The dictionary is not explicitly defined in the energy function but implicitly by the sparse coding (See (1)). We use the chain rule of differentiation to compute the gradient of ℓ^n with respect to the dictionary,

$$\frac{\partial \ell^n}{\partial \mathbf{D}} = \sum_{i\in\mathcal{V}} \left(\frac{\partial \ell^n}{\partial \mathbf{s}_i}\right)^\top \frac{\partial \mathbf{s}_i}{\partial \mathbf{D}}. \tag{15}$$

The difficulty of computing this gradient lies in that there is no explicit differentiation of sparse code $\mathbf{s}$ with respect to the dictionary $\mathbf{D}$. We overcome this problem by using implicit differentiation on the fixed point equation [29], [30]. We first establish the fixed point equation of (1),

$$\mathbf{D}^\top(\mathbf{D}\mathbf{s} - \mathbf{x}) = -\lambda \text{sign}(\mathbf{s}), \tag{16}$$

where sign($\mathbf{s}$) denotes the sign of $\mathbf{s}$ in a element-wise manner and sign(0) = 0. We compute the derivative of $\mathbf{D}$ on both sides of (16), and have

$$\frac{\partial \mathbf{s}_\Lambda}{\partial \mathbf{D}} = (\mathbf{D}_\Lambda^\top \mathbf{D}_\Lambda)^{-1}\left(\frac{\partial \mathbf{D}_\Lambda^\top \mathbf{x}}{\partial \mathbf{D}} - \frac{\partial \mathbf{D}_\Lambda^\top \mathbf{D}_\Lambda}{\partial \mathbf{D}}\right), \tag{17}$$

where Λ denotes the index set of non-zero codes of $\mathbf{s}$ and $\bar{\Lambda}$ is the index set of zero codes. To simplify the computation in (15), we introduce an vector of auxiliary variables $\mathbf{z}$ for each $\mathbf{s}$,

$$\mathbf{z}_{\bar{\Lambda}} = 0, \mathbf{z}_\Lambda = (\mathbf{D}_\Lambda^\top \mathbf{D}_\Lambda)^{-1} \cdot \frac{\partial \ell^n}{\partial \mathbf{s}_\Lambda}, \tag{18}$$

where $\partial \ell^n / \partial \mathbf{s}_\Lambda = (\mathbf{y}_i - \hat{\mathbf{y}}_i) \cdot \mathbf{w}_\Lambda$. In addition, we denote $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m]$. Therefore, the gradient of ℓ^n with respect to $\mathbf{D}$ is computed by

$$\frac{\partial \ell^n}{\partial \mathbf{D}} = -\mathbf{D}\mathbf{Z}\mathbf{S}^\top + (\mathbf{X} - \mathbf{D}\mathbf{S})\mathbf{Z}^\top. \tag{19}$$

The proposed joint learning algorithm is summarized in Algorithm 1.

5 Fixation Prediction

In free viewing scenarios, humans are usually attracted by familiar objects from daily life, such as people, animals, vehicles and household objects. Visual search of common objects serves as a top-down module of visual attention models which is complementary to the visual stimuli based bottom-up modules [8], [21]. In this section, we extend the proposed model to locate common objects and apply it to predict human fixations together with bottom-up saliency maps. Among all possible object categories, the PASCAL VOC dataset provides a collection of 20 common classes [16]:

- People: *Person*
- Animals: *Cat*, *Dog*, *Cow*, *Horse*, *Sheep* and *Bird*;
- Vehicles: *Car*, *Bus*, *Bicycle*, *Motorbike*, *Aeroplane*, *Boat* and *Train*;
- Household: *Chair*, *Sofa*, *Dining table*, *TV/monitor*, *Bottle* and *Potted plant*.

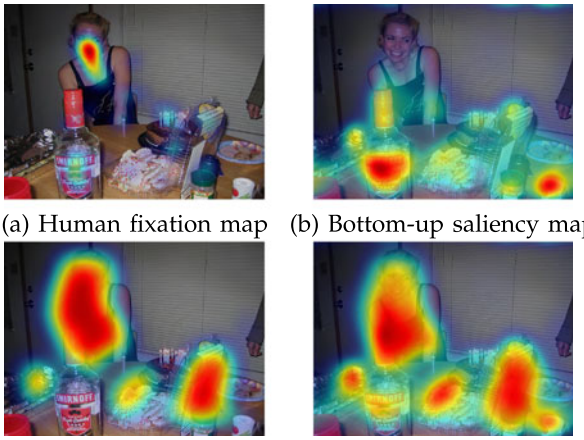
Algorithm 1. Joint CRF and Dictionary Learning

Input: $\mathcal{X}$ (training images) and $\mathcal{Y}$ (ground truth labels);
$\mathbf{D}^0$ (initial dictionary); $\mathbf{w}^0$ (initial CRF weight); λ (in (1));
T (number of cycles); γ (in (12)) ρ_0 (initial learning rate).
Output: $\hat{\mathbf{D}}$ and $\hat{\mathbf{w}}$.
Set $\hat{\mathbf{D}} = \mathbf{D}^0$, $\hat{\mathbf{w}} = \mathbf{w}^0$.
for $t = 1, \ldots, T$ **do**
  Permute training samples $(\mathcal{X}, \mathcal{Y})$
  for $n = 1, \ldots, N$ **do**
    Solve sparse coding $\mathbf{s}_i$ by (1), $\forall i \in \mathcal{V}$;
    Solve the most violated labeling $\hat{\mathbf{Y}}^{(n)}$ by (11);
    Update the CRF weight by (14):
      $\hat{\mathbf{w}} = \hat{\mathbf{w}} - \rho_t \frac{\partial \ell^n}{\partial \hat{\mathbf{w}}}$;
    Compute the auxiliary variables $\mathbf{z}_i$ by (18);
    Update the dictionary by (19):
      $\hat{\mathbf{D}} = \hat{\mathbf{D}} + \rho_t \frac{\partial \ell^n}{\partial \hat{\mathbf{D}}}$;
    Project the dictionary $\hat{\mathbf{D}}$ onto $\mathcal{D}$ by (7);
  end for
  Update the learning rate ρ: $\rho_t = \rho_0/t$
end for

We denote the common classes by a set of discrete labels $\mathbf{C} = 1, 2, \ldots, 20$. For a particular common object class $\mathbf{C} = c$, the proposed top-down saliency map predicts its presence on the patches of input image by $p(\mathbf{Y}|\mathbf{X}, \mathbf{C} = c)$. To locate generic common objects without target class assumption, we integrate class variable $\mathbf{C}$ over 20 common object classes to obtain category-independent top-down saliency by

$$p(\mathbf{Y}|\mathbf{X}) = \sum_{c=1}^{20} p(\mathbf{Y}|\mathbf{X}, \mathbf{C} = c)p(\mathbf{C} = c|\mathbf{X}). \tag{20}$$

The first term on the right side of (20) is the top-down saliency map for common object class $\mathbf{C} = c$ introduced in previous sections while the second term $p(\mathbf{C} = c|\mathbf{X})$ is the probabilistic output of object categorization models that provides global modulation for a particular object class. In other words, the proposed top-down saliency map and object categorization captures the "where" and "what"' components for object recognition, respectively. Technically, any object classification models will suffice. Given an input image $\mathbf{X}$, we calculate its top-down saliency map for common objects $\mathbf{O}$ from $p(\mathbf{Y} = 1|\mathbf{X})$ using the normalization method in Section 3.3 and its bottom-up saliency map $\mathbf{U}$ using state-of-the-art algorithms [3], [5]. We use a binary



(a) Human fixation map (b) Bottom-up saliency map

(c) Top-down saliency map (d) Combined map

Fig. 2. Comparison of bottom-up and top-down saliency maps for human fixation prediction. Warmer color (from red to blue) indicates higher saliency value.

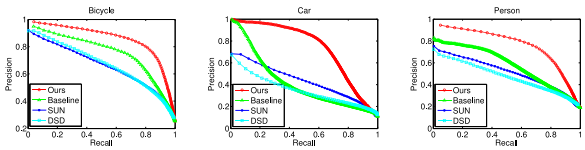


Fig. 3. Patch-based precision-recall curves on Graz-02 dataset.

variable $\mathbf{f}_i$ to represent human fixation at pixel i and predict eye fixation by

$$p(\mathbf{f}_i) = \alpha \mathbf{u}_i + (1-\alpha)\mathbf{o}_i, \tag{21}$$

TABLE 1
Precision Rates (%) at EER on the Graz-02 Dataset

	Bicycle	Car	Person
DSD [17]	62.5	37.6	48.2
SUN [8]	61.9	45.7	52.2
Baseline, $k = 512, \lambda = 0.15$	71.9	39.3	56.8
Joint, $k = 256, \lambda = 0.15$	73.3	57.5	64.2
Joint, $k = 512, \lambda = 0.15$	**80.1**	**68.6**	**72.4**
Joint, $k = 512, \lambda = 0.30$	73.5	66.6	69.6

where $\mathbf{u}_i \in \mathbf{U}$ and $\mathbf{o}_i \in \mathbf{O}$ are the bottom-up and top-down saliency values at pixel i and α is the tradeoff constant value that usually is set to 0.5. Fig. 2 presents an example of fixation prediction by combining bottom-up and top-down saliency maps.

6 Experiments

We evaluate the proposed top-down saliency algorithm in the context of object localization and fixation prediction. In the Graz-02 experiments, we compare the proposed model with two state-of-the-art top-down saliency map algorithms and show its performance on object localization. In the PASCAL VOC experiments, we present multiscale top-down saliency maps for 20 object classes and analyze their cross-category performance. We then apply our top-down saliency maps to fixation prediction tasks using the PASCAL_S dataset [36].

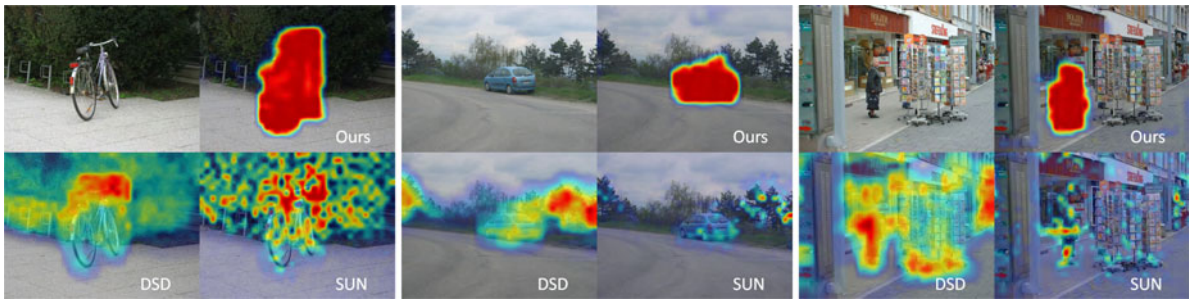
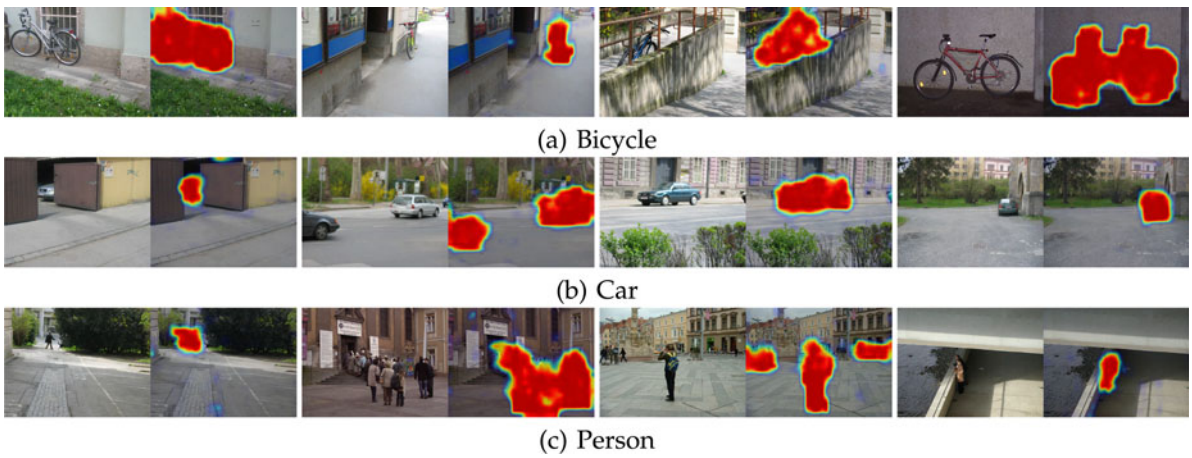


Fig. 4. Top-down saliency maps generated by the proposed, DSD and SUN models.



(a) Bicycle

(b) Car

(c) Person

Fig. 5. Saliency maps of bicycle, car and person categories from the Graz-02 dataset generated by the proposed algorithm. In each panel, we present the original image and the saliency map, respectively. Overall, the proposed saliency maps are able to locate objects with large viewpoint changes, scale variations and heavy occlusions.

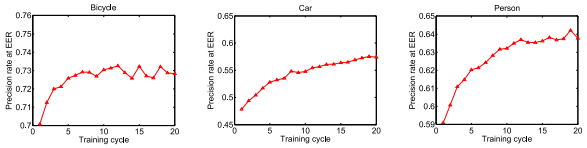


Fig. 6. Performance gain with training cycles. The dictionary size $k = 256$ and the sparsity regularization term $\lambda = 0.15$.

6.1 Graz-02

The Graz-02 dataset contains three categories (bicycles, cars and persons) and one background class. Each category has 300 images of size 640×480 pixels and the corresponding pixel-based foreground/background annotations. We choose this dataset because all of three categories contain real-world images with large intra-class variations, occlusions and background clutters. The task is to evaluate the performance of top-down saliency maps to localize target objects against the background.

Implementations. We sample image patches of 64×64 pixels by shifting 16 pixels and collect 999 patches on a 27×37 grid for each image. The SIFT descriptors [37] are extracted from each image patch to represent the object appearance. We label a patch as positive if at least one quarter of its total pixels are foreground, and obtain a binary patch-based saliency mask from the original pixel-based annotation of each image. For each category, we use 150 odd-numbered foreground images and 150 odd-numbered background images as the training set, and the remaining 150 foreground and 150 background images as the test set.

To train the proposed saliency model by Algorithm 1, we need to initialize the dictionary and the CRF model. We collect all these SIFT descriptors from the training set and use the K-means algorithm to initialize the dictionary $\mathbf{D}^{(0)}$. After evaluating the latent variables by sparse coding, we initialize the CRF node energy weight $\mathbf{w}_1^{(0)}$ by training a linear SVM on the sparse codes and the corresponding saliency labels. All the models are trained with 20 cycles.

Parameter settings. There are two important parameters in our model. One is the number of visual words (atoms) K in the dictionary, which controls the capacity of modeling appearance variations. Although it is usually more effective to model object appearance with a larger dictionary, it is more difficult and time-consuming to achieve this as more training examples are required. In our experiments, we train saliency map models with 256 or 512 visual words. The other parameter is the sparsity regularization term λ defined in (1). The greater the λ is, the more sparse the latent codes are and the fewer visual words are selected to represent an image patch. We use two values, 0.15 and 0.30, for λ in the experiments. In Algorithm 1, we set the initial learning rate $\rho_0 = 1e-3$ and the weight regularization term $\gamma = 1e-5$.

Comparisons with state-of-the-art methods [17], [38]. We compare our model with two state-of-the-art top-down saliency algorithms [8], [17] by using our own implementations. To demonstrate the effectiveness of joint CRF and dictionary learning, we also construct a baseline model by switching off the dictionary update module. For the discriminant saliency detection algorithm (DSD) [17], we first construct a dictionary based on the Discrete Cosine Transform (DCT) with 256 filters of size 64×64, and then select 100 salient features with largest mutual information. For the saliency using natural statistics algorithm (SUN) [8], we first reduce the dimension of the image patches by Principle Component Analysis (PCA) and then learn 724 filters by Independent Component Analysis (ICA) from the training data. By using the ICA filter responses as features, a linear SVM is trained to compute the saliency values of patches.

All the models (ours, baseline, DSD, and SUN) are evaluated by patch-based precision-recall rates on the test set of each category. Fig. 3 shows the precision-recall curves for three object categories, respectively. Overall, the proposed saliency map algorithm performs favorably against the state-of-the-art methods. Furthermore, the results demonstrate the importance of dictionary update in the proposed algorithm.

In Table 1, we present the results using different parameters (k, λ) of all the models in terms of precision rates at equal error rates (EER where precision is equal to recall). The best results are obtained by our model with the parameters $k = 512, \lambda = 0.15$. The results also show substantial improvements of our models over the baseline and other algorithms. The DSD algorithm selects salient features based on image statistics that usually have limited ability of suppressing background. In general, the DSD method generates high recall but low precision rates. The SUN algorithm performs better than the DSD method which can be attributed to the use of strong classifiers. Without considering local contexts, the SUN algorithm tends to produce noisy saliency maps. Our models are able to produce clear saliency maps when target objects appear in different viewpoints and scales with substantial occlusions. A saliency map of an image has the size of its patch grid, i.e., 27×37. We upsample the original saliency map to the size of image by bilinear interpolation. Fig. 4 shows the saliency maps generated by the DSD, SUN and proposed algorithms. Note that the proposed saliency algorithm is able to locate heavily occluded objects (e.g., bicycle and cars) whereas

TABLE 2
PASCAL VOC 2007 Localization Results

	aero	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dtable	dog	horse	mbike	person	plant	sheep	sofa	train	tv	average
# of training images	238	243	330	181	244	186	713	337	445	141	200	421	287	245	2008	245	96	229	261	256	365
# of test images	204	239	282	172	212	174	721	322	417	127	190	418	274	222	2007	224	97	223	259	229	351
Baseline	79.7	76.5	70.6	73.7	47.9	74.4	77.8	74.1	52.3	78.3	76.4	73.9	77.6	77.3	76.1	62.0	79.2	75.4	75.9	59.7	71.9
Joint	80.2	79.3	72.6	74.6	57.7	79.3	81.2	75.2	58.0	79.8	77.4	75.7	81.0	79.4	78.6	66.6	79.9	77.1	78.4	70.6	75.1

In each column, we present the category, the numbers of training and test images, the precision rates at EER by the baseline and joint learning algorithms, respectively.

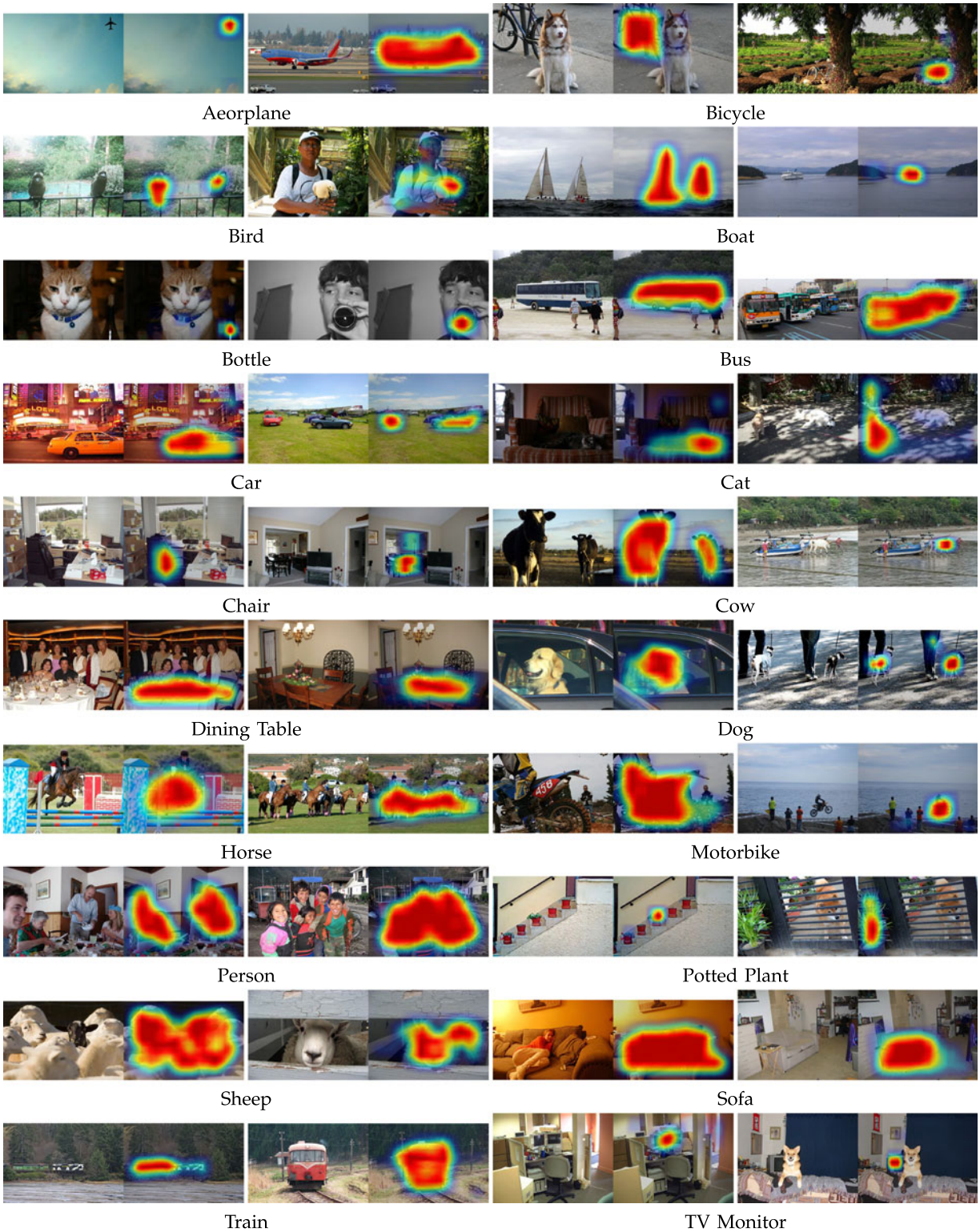


Fig. 7. Within-category saliency detection results. We present representative saliency maps from 20 categories in a 10×2 table and each cell includes two test cases where the original image is on the left while the saliency map is on the right. The lowest to highest saliency measures are shown in color from blue to red (the saliency maps are best viewed in color).

state-of-the-art object detection methods are not expected to perform well in such cases.

Effect of dictionary update. Our saliency model jointly learns CRF weights and dictionary from the training examples by gradient updates (Algorithm 1). We are interested in how the dictionary update helps improve the model performance. Thus, we evaluate the CRF weights and the dictionary at each training cycle on the test set. Fig. 6 presents the precision rates at EER of each cycle. As shown in the figure, the performance improves substantially in the first several cycles and converge after 10 cycles. The stochastic nature of our learning algorithm

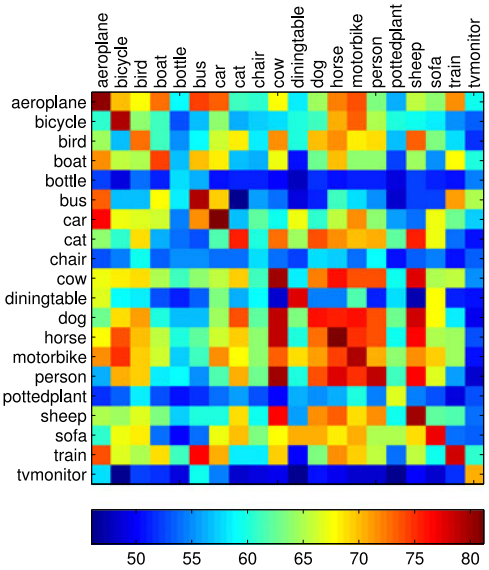


Fig. 8. Confusion matrix for cross-category saliency maps. The red dot denotes the class with high saliency precision while the blue dot denotes the class with low saliency precision.

results in some performance perturbation at some cycles. The results show that dictionary update significantly improves the model performance.

6.2 PASCAL VOC

The PASCAL VOC 2007 dataset consists of 5,011 images for training and validation, and 4,952 images for tests from 20 object categories and one background class. All the images are annotated with bounding boxes while segmentation annotations are available for 632 images. This dataset is more challenging for top-down visual saliency because objects from different categorizes may appear in the same image with cluttered background. We first evaluate our saliency model for localization and then apply the saliency maps for fixation prediction.

6.2.1 Object Localization

The saliency models are trained in a class-wise manner. For each class, we only use the positive images including target objects for training. This training strategy conforms with the aim of top-down visual saliency to discriminate target

Fig. 9. Affinity of 20 categories in a two-dimensional space by their Laplacian eigenmap. The red dot denotes the class with high saliency precision while the blue dot denotes the class with low saliency precision.

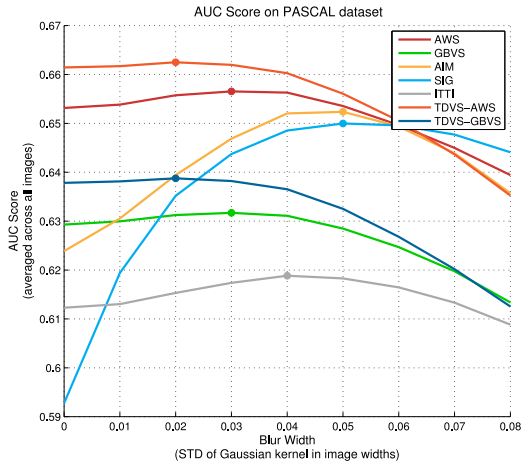


Fig. 10. Comparison of AUC scores of various state-of-the-art saliency maps (indicated by different colors) for fixation prediction on the PASCAL_S dataset. The dots on the curves show the best AUC scores are obtained using Gaussian blur kernels.

objects from their surroundings. When applied to negative images, our model is expected to highlight most target-like (salient) regions. This differentiates our saliency model from object detection models that are trained to suppress all the possible false positives. In the training phase, we extract image patches at three scales 48×48, 64×64 and 80×80 from a denser grid of every 4 pixels. We train saliency models at three scales separately and then combine the saliency maps in the test phase. The training process requires computing sparse coding of each patch for numerous iterations. We use a similar method with the Graz-02 experiments to generate saliency masks from labeled segmentations. For those images without labeled segmentations, we create masks by measuring whether the sampled patches fall into target bounding boxes. We set the dictionary size to $K = 512$ and $\lambda = 0.15$ to train the models with 10 cycles for all classes. To demonstrate the effectiveness of joint dictionary and CRF learning, we also train the models with baseline algorithms without dictionary update.

Within-category results. For each class, we first evaluate the learned model with the corresponding positive images. We measure the performance by the precision rates at EER and present the results from the baseline and joint learning algorithms in Table 2. The proposed joint learning algorithm consistently outperforms the baseline method which demonstrates the merits of dictionary update. Fig. 7 shows representative saliency maps from each category. These results demonstrate that the proposed model is able to handle large scale and viewpoint changes (e.g., aeroplanes), significant lighting variations (e.g., cats), and partial occlusions and articulations (e.g., people), However, our model is more likely to locate objects with rich textures due to the use of the adopted SIFT-based patch representation. Table 2 and Fig. 7 show that our model performs better in the classes of aeroplanes, bicycles and horses than in the classes of bottles and chairs. This is likely because aeroplanes, bicycles and horses are easier to identify from their shapes. In addition, we observe that some saliency models tend to locate other objects that co-occur with the targets in the training images. For example, the saliency model for dogs tend to highlight people and cats as well (which can be

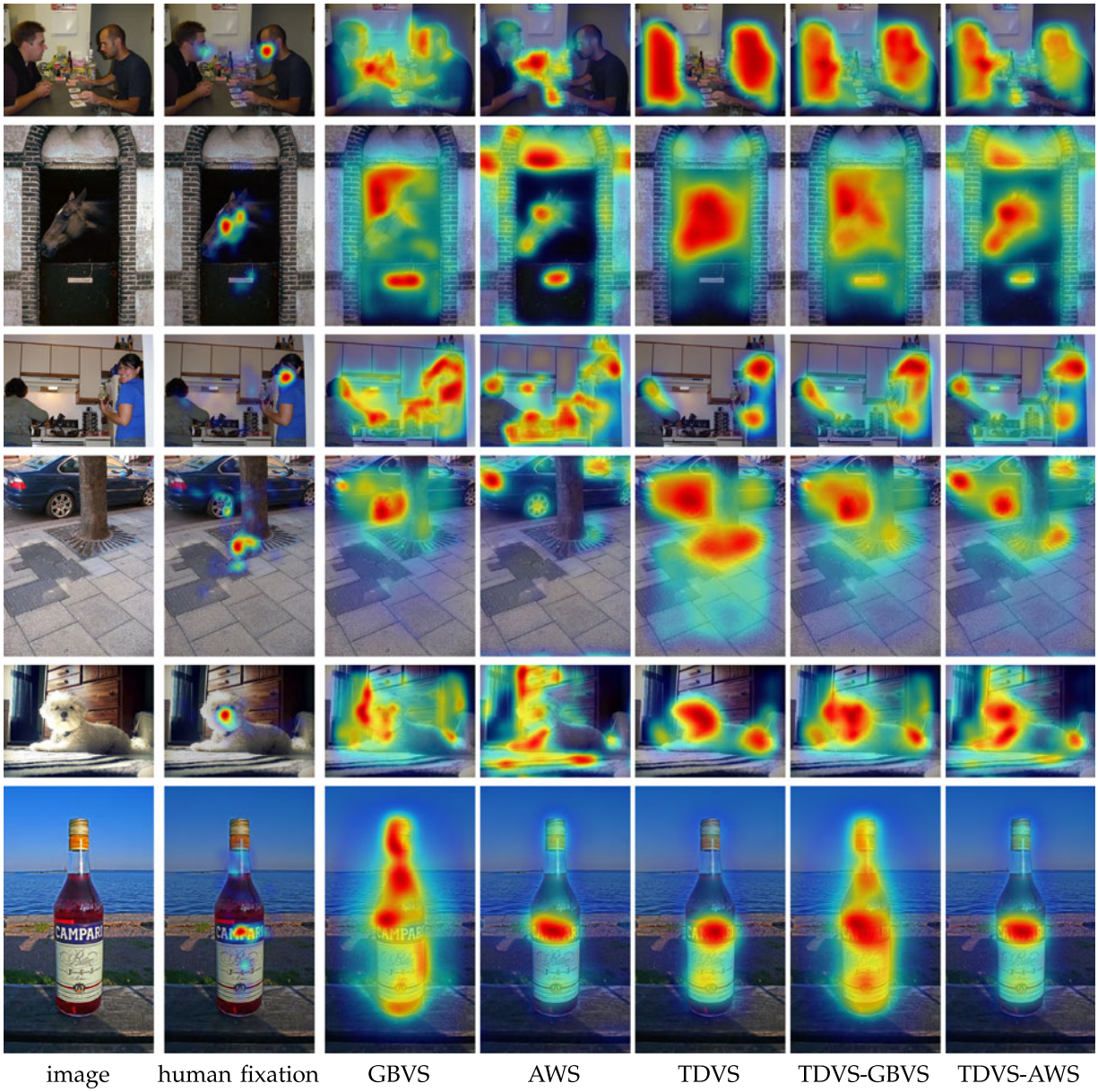


Fig. 11. Qualitative fixation prediction results. In each row, we present an input image, the ground truth fixation map, two bottom-up saliency maps (GBVS and AWS), our category-independent top-down saliency map (TDVS) and combined saliency maps (TDVS-GBVS and TDVS-AWS). The lowest to highest saliency measures are shown in color from blue to red (the saliency maps are best viewed in color).

attributed by the facts that these objects tend to appear in the same training images). This fact motivates us to investigate the performance of our model on the negative images. More specifically, we are interested in evaluating the saliency models across categories.

Cross-category results. We apply the saliency model for one category to the test sets of all the other 19 categories, and compute the precision rates at EER based on the ground truth saliency masks. The precision rates are summarized in the confusion matrix $C(\cdot,\cdot)$ shown in Fig. 8, where the entry $C(i,j)$ represents the precision rate of saliency model i on the test set of category j. These results indicate the ability of the saliency model of one class to highlight the object regions of another class. It is of interest to observe that some saliency models perform quite well in the test sets of particular classes, e.g., dog model in cats images and cow model in sheep images. This can be explained by the mutual saliency between two object classes due to patch-level appearance similarity.

In order to better analyze mutual saliency among categories, we show the confusion matrix of Fig. 8 in a two-dimensional embedded space using the Laplacian eigenmaps [39]. The cross-category precision in the entry $C(i,j)$ represents how well the model of class-i performs in the images of class-j, and likewise for $C(j,i)$. The average precision, $(C(i,j)+C(j,i))/2$, thus represents the affinity between class-i and class-j, and the matrix $A=(C+C^{\top})/2$ denotes the affinity matrix of 20 classes. We extract the first two eigenvectors of affinity matrix A as two-dimensional embedded coordinates, which are depicted in Fig. 9. The embedded results of the affinity matrix can be split into three clusters:

1) person, dog, cat, cow, sheep, horse, bird, motorbike, bicycle and sofa;

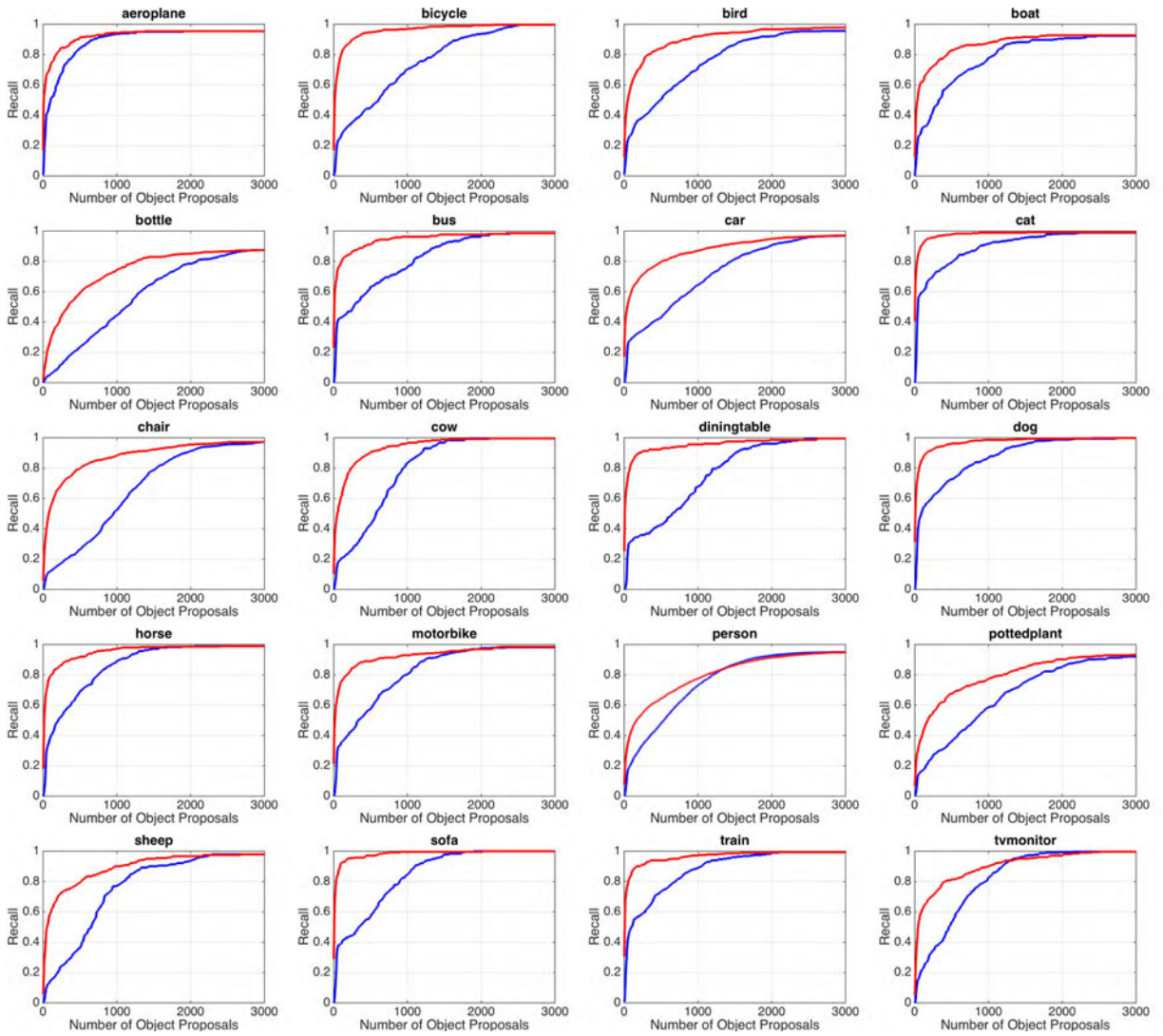


Fig. 12. Evaluating top-down saliency maps by recall rates of object proposals. In each panel, the blue curve denotes the recall rates given by the bounding box proposals drawn from selective search [13] while the red curve denotes the recall rates of those proposals ranked by their top-down saliency values.

2) aeroplane, car, boat, train and bus;
3) pottedplant, chair, diningtable, bottle and tvmonitor.

There are two factors that support this mutual saliency relationship, i.e., feature sharing and object co-occurrence. All the animals share similar part configurations (e.g., head, body and legs) such that the classes of person, dog, cat, cow, sheep, horse and bird follow into the first cluster. In the second cluster, all the classes of aeroplane, car, boat, train and bus belong to large vehicles which consist of wheels, windows and other rigid structures. Interestingly, the motorbike and bicycle classes also fall into the first cluster. This can be attributed to the fact that motorbikes, bicycles and sofa usually co-occur with people in the training images. In this dataset, many images include people riding bicycles, people riding motorbikes or people lying in the sofa. As most saliency masks are obtained from bounding box annotations, it is inevitable that patches of concurrent objects (bicycles, motorbikes and sofa) are considered as positive examples and vice versa. Many classes in the third cluster have low precisions in the test sets. It is thus less interesting to investigate mutual saliency among these classes. However, objects from these classes (e.g., potted plant, chair, dining table, bottle and tv monitor) appear frequently in households.

Feature sharing [40] and object co-occurrence [41] are important image structures for object recognition. Our experimental results show that such properties can be obtained via clustering on saliency maps generated by the proposed algorithm.

6.2.2 *Fixation Prediction*

We present experimental results for fixation prediction based on the method introduced in Section 5 and the PASCAL_S dataset [36]. The PASCAL_S dataset consists of 850 images from the validation set of PASCAL VOC 2010 and fixation data is collected from eight subjects using Eyelink 1,000 eye-tracker. Each image is presented to subjects for 2 seconds in the free viewing scenarios. Thorough performance evaluation of state-of-the-art saliency algorithms for fixation prediction are presented [36]. We train image classifiers on the training set of PASCAL VOC 2007. For each image, we extract features from the seventh layer of a deep convolutional network pre-trained

on the ImageNet dataset [18], [42], and train classifiers for 20 object classes using linear SVM classifiers [43]. We compute the category-independent top-down saliency map $\mathbf{O} = p(\mathbf{Y} = 1|\mathbf{X})$ (20) for each image in the PASCAL_S dataset using the the saliency models in previous localization experiments $p(\mathbf{Y} = 1|\mathbf{X}, \mathbf{C} = c)$ and the probabilistic output of SVM classifiers $p(\mathbf{C} = c|\mathbf{X})$. We combine our category-independent top-down saliency maps (TDVS) with two state-of-the-art bottom-up saliency maps, AWS [5] and GBVS [3] as our predictions using (21), and the results are denoted by TDVS-AWS and TDVS-GBVS, respectively. We compare with other representative fixation prediction algorithms, ITTI [1], AIM [2] and SIG [44]. All the saliency maps are blurred with Gaussian kernels by varying the bandwidth. We compute the AUC (area under ROC curve) scores for compared saliency maps with different Gaussian blur kernels by using the code provided by [36]. The results in Fig. 10 show that combing top-down and bottom-up saliency maps is able to improve the performance for human fixation prediction (See the improvements of TDVS-GBVS over GBVS and TDVS-AWS over AWS). Some qualitative results are presented in Fig. 11. Overall, our top-down saliency maps generates high recall of object regions that help predict human fixations when bottom-up saliency gets distracted by cluttered backgrounds.

6.2.3 Prioritizing Object Proposals

State-of-the-art object detectors [11] need to evaluate object proposals on deep convolutional networks on thousands [13]. Since many of them are simply background patches or irrelevant objects, the learned category-specific top-down saliency maps can be used to prune object proposals before feeding into deep convolutional networks. Specifically, we calculate a saliency score for each object proposal box, which provides category-specific rankings for object proposals as shown in Fig. 12. In most classes, the method based on saliency ranking achieves above the recall rate of 90 percent with 1,000 proposals (about half of total proposals typically used in the state-of-the-art methods) per image. Note that it takes less than 2 seconds to compute top-down saliency maps for three scales (including SIFT feature extraction, sparse coding and BP inference) for a 256×256 image using unoptimized MATLAB code on a desktop computer with an Intel i7 processor.

7 Concluding Remarks

We present a novel top-down visual saliency model via joint CRF and dictionary learning. Compared with computing saliency values individually on each patch by Gao et al. [17], and Kanan et al. [8], our saliency map is generated by considering the label consistency via the proposed layered CRF model. Our model thus produces clear saliency maps by leveraging local context information. We observe that significant improvements can be achieved by updating the dictionary under supervision of the proposed CRF model. The learned top-down saliency maps are used to prioritize object proposals for object detection. We extend our model to category-independent top-down saliency and show that it provides complementary information to bottom-up saliency for improving fixation prediction.

Acknowledgments

The work is supported in part by NSF CAREER Grant #1149783 and NSF IIS Grant #1152576.

References

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene anaysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[2] N. D. B. Bruce and J. K. Tsotsos, "Saliency based on information maximization," in *Proc. Int. Conf. Neural Inform. Process. Syst.*, 2005, pp. 155–162.

[3] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Neural Inform. Process. Syst.*, 2006, pp. 545–552.

[4] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recogn.*, 2007, pp. 1–8.

[5] A. Garcia-Diaz, V. Leboran, X. R. Fdez-Vidal, and X. M. Pardo, "On the relationship between optical variability, visual saliency, and eye fixations: A computational approach," *J. Vis.*, vol. 12, no. 6, pp. 1–22, 2012.

[6] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[7] S. Chikkerur, T. Serrea, C. Tana, and T. Poggio, "What and where: A Bayesian inference theory of attention," *Vis. Res.*, vol. 50, pp. 2233–2247, 2010.

[8] C. Kanan, M. H. Tong, L. Zhang, and G. W. Cottrell, "Sun: Top-down saliency using natural statistics," *Vis. Cognition*, vol. 17, no. 8, pp. 979–1003, 2009.

[9] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 65–81, Feb. 2004.

[10] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu, "Semantic segmentation with second-order pooling," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 430–443.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2015.

[12] J. Carreira and C. Sminchisescu, "Constrained parametric min-cuts for automatic object segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recogn.*, 2010, pp. 3241–3248.

[13] K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders, "Segmentation as selective search for object recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1879–1886.

[14] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2189–2202, Nov. 2012.

[15] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer, "Generic object recognition with boosting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 3, pp. 416–431, Mar. 2006.

[16] M. Everingham, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[17] D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 6, pp. 989–1005, Jun. 2009.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Int. Conf. Adv. Neural Inform. Process. Syst.*, 2012, pp. 1106–1114.

[19] A. Torralba, A. Oliva, M. Castelhano, and J. M. Henderso, "Contextual guidance of attention in natural scenes: The role of global features on object search," *Psychological Rev.*, vol. 113, no. 10, pp. 766–786, 2006.

[20] S. Mathe and C. Sminchisescu, "Action from still image dataset and inverse optimal control to learn task specific visual scanpaths," in *Proc. Int. Conf. Neural Inform. Process. Syst.*, 2013, pp. 1923–1931.

[21] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. 12th Int. Conf. Comput. Vis.*, 2009, pp. 2106–2113.

[22] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao, "Predicting human gaze beyond pixels," *J. Vis.*, vol. 14, no. 1, pp. 1–20, 2014.

[23] L. Bertelli, T. Yu, D. Vu, and B. Gokturk, "Kernelized structural SVM learning for supervised object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 2153–2160.
[24] B. Fulkerson, A. Vedaldi, and S. Soatto, "Class segmentation and object localization with superpixel neighborhoods," in *Proc. 12th Int. Conf. Comput. Vis.*, 2009, pp. 670–677.
[25] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1848–1853, Oct. 2007.
[26] A. Jain, L. Zappella, P. McClure, and R. Vidal, "Visual dictionary learning for joint object categorization and segmentation," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 718–731.
[27] L. Tao, F. Porikli, and R. Vidal, "Sparse dictionaries for semantic segmentation," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 549–564.
[28] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," in *Proc. Neural Inform. Process. Syst.*, 2008, p. 15.
[29] J. Yang, K. Yu, and T. Huang, "Supervised translation-invariant sparse coding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 3517–3524.
[30] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 791–804, Apr. 2011.
[31] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent K-SVD," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 1697–1704.
[32] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Proc. Neural Inform. Process. Syst.*, 2006, pp. 801–808.
[33] M. Szummer, P. Kohli, and D. Hoiem, "Learning CRFS using graph cuts," in *Proc. 10th Eur. Conf. Comput. Vis.*, 2008, pp. 582–595.
[34] T. Joachims, T. Finley, and C.-N. Yu, "Cutting-plane training of structural SVMS," *Mach. Learn.*, vol. 77, no. 1, pp. 27–59, 2009.
[35] C.-N. J. Yu and T. Joachims, "Learning structural SVMS with latent variables," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 1169–1176.
[36] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 280–287.
[37] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[38] C. Kanan and G. Cottrell, "Robust classification of objects, faces, and flowers using natural image statistics," in *Proc. Int. Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 2472–2479.
[39] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. Adv. Neural Inform. Process. Syst.*, 2001, pp. 585–591.
[40] A. Torralba, K. P. Murphy, and W. T. Freeman, "Sharing visual features for multiclass and multiview object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 5, pp. 854–869, May 2007.
[41] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr, "Graph cut based inference with co-occurrence statistics," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 239–253.
[42] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multi., arXiv:1408.5093*, 2014, pp. 675—678.
[43] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, 2001, Art. no. 27.
[44] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 194–201, Jan. 2011.

Jimei Yang received the BS degree in electrical engineering and information science from the China Agricultural University and the MEng degree in pattern recognition and intelligent systems from the University of Science and Technology of China in 2006 and 2009, respectively. He received the PhD degree in computer science from the University of California, Merced. He was a visiting PhD student in Artificial Intelligence Lab, University of Michigan, Ann Arbor, in 2015. He is currently a research scientist at Adobe Research, San Jose. From June 2007 to June 2009, he worked as a research assistant at Institute of Automation, Chinese Academy of Sciences.

Ming-Hsuan Yang received the PhD degree in computer science from the University of Illinois, Urbana-Champaign in 2000. He is currently an associate professor in electrical engineering and computer science at the University of California, Merced. Prior to joining UC Merced in 2008, he was a senior research scientist at the Honda Research Institute working on vision problems related to humanoid robots. He served as an associate editor of the *IEEE Transactions on Pattern Analysis and Machine Intelligence* from 2007 to 2011, and is an associate editor of the *International Journal of Computer Vision*, *Image and Vision Computing* and *Journal of Artificial Intelligence Research*. He received the NSF CAREER Award in 2012, the Senate Award for Distinguished Early Career Research at UC Merced in 2011, and the Google Faculty Award in 2009. He is a senior member of the IEEE and the ACM.