Top down saliency estimation via superpixel-based discriminative dictionaries

Aysun Kocak
aysunkocak@cs.hacettepe.edu.tr
Kemal Cizmeciler
kemalcizmeci@gmail.com
Aykut Erdem
aykut@cs.hacettepe.edu.tr
Erkut Erdem
erkut@cs.hacettepe.edu.tr

Computer Vision Lab
Department of Computer Engineering
Hacettepe University
Ankara, Turkey

Abstract

Predicting where humans look in images has gained significant popularity in recent years. In this work, we present a novel method for learning top-down visual saliency, which is well-suited to locate objects of interest in complex scenes. During training, we jointly learn a superpixel based class-specific dictionary and a Conditional Random Field (CRF). While using such a discriminative dictionary helps to distinguish target objects from the background, performing the computations at the superpixel level allows us to improve accuracy of object localizations. Experimental results on the Graz-02 and PASCAL VOC 2007 datasets show that the proposed approach is able to achieve state-of-the-art results and provides much better saliency maps.

1 Introduction

Predicting salient parts of an image that attract attention has recently gained a huge interest in computer vision. This is particularly because such a prediction allows to filter out irrelevant information within an image and to focus on the bits that are really important. In that regard, visual saliency has been used to improve the performance of different computer vision tasks, including image retargeting [1], video compression [18], video summarization [21], object detection [25], object recognition [31, 37], object tracking [7], scene classification [33].

Most works on visual saliency concentrate on predicting human eye fixations (see [5] for a recent survey). However, there is an increasing number of studies that aim at the alternative task of detecting salient objects [3, 6, 20, 23, 24, 27, 38, 39]. These recent methods can be categorized into two groups as bottom-up (e.g. [23, 27, 38]) and top-down (e.g. [3, 20, 24, 39]) approaches according to how they define salient objects. Bottom-up models mostly rely on low-level cues such as intensity, color, texture, etc. and try to localize objects which show distinct characteristics from their surroundings. On the other hand, top-down approaches are task-oriented and look for a target object from a specific category. Hence, they employ appearance characteristics of the object of interest.

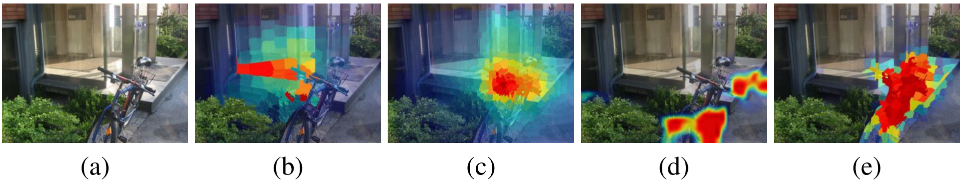

Figure 1: Bottom-up vs. top-down saliency. (a) Input image, and the results of (b) a bottom-up salient object detection method [38], (c) the objectness map generated by [4], (d) the top-down saliency model of [39] and (e) our approach. Bottom-up model and generic objectness approach do not capture the object of interest (the bike in this example) due to lack of prior knowledge. The top-down saliency model of [39] partly captures the bike in the input image. Our saliency model, on the other hand, provides a considerably better localization than [39].

In this paper, we propose a novel method for top-down saliency estimation to generate category-specific saliency maps. Our approach is inspired in part by the recent dictionary-based top-down saliency approaches [20, 39] and the new superpixel-based bottom-up salient object detection methods [23, 27, 38]. Specifically, both [39] and [20] learn sparse coding based discriminative dictionaries within a CRF framework to model appearance properties of object classes. This results in a better representation of target objects and thus better saliency maps. However, the shortcoming of those approaches is that they are patch-based, i.e they estimate patchwise saliency scores and in turn provide very coarse localizations of the target objects. Towards improving object localizations and obtaining better saliency maps, we suggest a more effective way to represent objects and a more accurate inference mechanism. As shown in Fig. 1, the proposed method provides considerably better top-down saliency maps.

We describe our contributions and their relation to closely related previous work in detail in the next two sections – but to summarize, they are as follows: (1) We propose to learn discriminative dictionaries at a superpixel-based level, instead of a patch-based one. We suggest to use the so-called sigma points descriptor [17] to represent superpixels in terms of first and second-order feature statistics. These allow us to encode the visual characteristics of the targets objects in a compact way. (2) Generic bottom-up information about objects has shown to play an important role in object detection (e.g. [4, 28, 36]). We also propose to include such generic objectness cues into our framework to further boost the performance. This gives much better localizations of the salient objects in the images than the previous related models. We evaluate our model on the Graz-02 and PASCAL VOC 2007 benchmark datasets and demonstrate substantial improvements in the estimated saliency maps upon current state-of-the-art methods.

2 Related Work

Salient object detection methods can be categorized as bottom-up and top-down approaches. Another relevant methods is the generic objectness models. We briefly review them below.

Bottom-up salient object detection. The goal of bottom-up salient object detection is to identify the prominent objects in an image that attract attention under a free-viewing condition. They mainly differ from each other in how they define and estimate the saliency by means of different contrast measures. Recent examples include [23, 27, 38] (for a comprehensive survey, refer to [6]). Perazzi *et al.* [27] segment an image into superpixels and estimate saliency over them by considering two contrast measures which respectively depends on their uniqueness and color distribution. Margolin *et al.* [23] propose a combined framework that involve patch and superpixel based strategies, which are respectively based on pattern and color rarity. Yang *et al.* [38] detect salient regions in images by solving a graph-based manifold ranking problem, which is defined through some local grouping cues and background priors. These bottom-up methods show good performance on the existing (bottom-up) salient object detection datasets. However, the priors they depend on are generic in nature (do not use class knowledge), and thus they generate unsatisfactory results as illustrated for the *look for a bike* task in Fig. 1(b).

Generic objectness models. A line of research in object recognition which shares many similarities with bottom-up salient object detection focuses on localizing all objects in an image in a class-independent manner [4, 8, 9, 13, 22, 29, 36]. These generic objectness models integrate multiple cues, which all measure the distinctive yet generic characteristics of objects like having a closed boundary, being apparently different from their surroundings, etc., to generate a small number of object proposals in terms of either windows [4, 9, 22] or segments [8, 13, 29, 36] without employing any category specific knowledge. However, the fact that they still mostly rely on bottom-up cues makes them harder to locate a specific object of interest (Fig. 1(c)).

Top-down salient object detection. In computer vision literature, there are numerous approaches to computational models of bottom-up saliency. Not much work, however, has focused on task-oriented top-down estimation of visual saliency [12, 19, 20, 34, 39]. One early model for search tasks on real-world images is Torralba *et al.*'s contextual guidence model [34] which is derived under a Bayesian framework and combines low-level saliency and scene context. Later, object appearance cues are also incorporated to this model by Ehinger *et al.* [12] and Kanan *et al.* [19]. Compared to these former studies, the closest works to ours are [20, 39]. These models jointly learn category-specific dictionaries and the classifier weights within a CRF framework, and then use this graphical model to estimate top-down saliency maps of images. Lastly, there are also some studies which computes weight maps of images for tasks such as object recognition [10, 24], object segmentation [3, 16] and image classification [32], which are intrinsically connected to top-down saliency.

3 Our Approach

We take joint CRF and dictionary learning strategy from [20, 39], and suggest several novel improvements that allow us to obtain more accurate predictions. Specifically, our approach employs discriminative dictionaries but instead of densely extracting patches from images and carrying out the saliency estimation over them, we use superpixels and associate each one of them with a saliency score as recently done in most of the bottom-up salient object detection methods. We propose to use the sigma points descriptor [17] which allows us to represent superpixels in a more effective manner by encoding statistical relationships among simple visual features. Furthermore, to boost the performance, we additionally incorporate bottom-up generic objectness cues into our framework. The effects of these developments are clearly visible in Figure 1(d) and (e) that we extract the target object more accurately.

An overview of our framework is summarized in Fig. 2. Given a set of training images containing object level annotations, we first segment the images into superpixels and represent them with the sigma points descriptor [17] (Sec. 3.1). Additionally, we extract

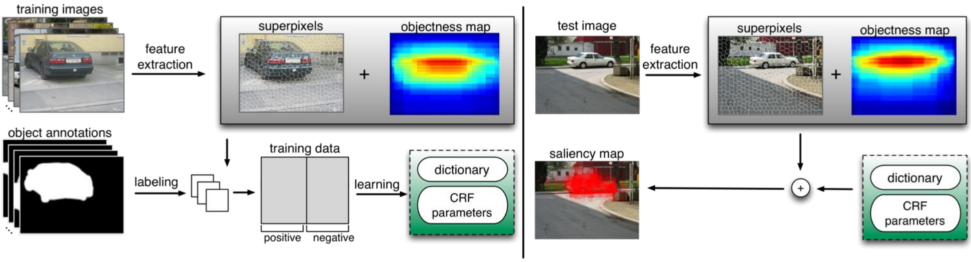

Figure 2: System overview. Given a training set of positive and negative images for an image category together with the salient/non-salient regions we jointly learn a superpixel-based discriminative dictionary as well as the CRF parameters. For a test image, the inference is carried out via the learned dictionary and by taking into account the objectness map.

objectness maps of these images. For each object category, we then jointly learn a dictionary and a CRF, which leads to a discriminative model that better distinguishes target objects from the background. When given a test image and a search task, we compute sparse codes of superpixels with dictionaries learned from data, estimate the objectness map and use the CRF model to infer saliency scores (Sec. 3.2).

3.1 Superpixel Representation

We segment the images into superpixels using the SLIC method [2][1]. After generating superpixels, we represent each one by means of the the first and the second order statistics of simple visual features. Note that most of the superpixel-based salient object detection methods represent superpixels by means of mean color features or color feature histograms. However, those type of representations fail to capture relationships between different feature dimensions. Hence, in this work, we propose to use the sigma points descriptor [17], which is based on region covariances [35] and has been previously explored for bottom-up saliency estimation in [11, 14].

Let $F(x,y) = \phi(I,x,y)$ denote the feature image extracted from an image I with ϕ denoting a mapping function that extracts an d-dimensional feature vector (such as constructed from intensity, color, orientation, pixel coordinates, etc.) from each pixel $i \in I$. Then, a region R inside F can be represented with a $d \times d$ covariance matrix $\mathbf{C}_R$:

$$\mathbf{C}_R = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{z}_i - \boldsymbol{\mu})(\mathbf{z}_i - \boldsymbol{\mu})^T \tag{1}$$

where $\mathbf{z}_{i=1\ldots n}$ denotes the d-dimensional feature vectors inside R and μ is the mean of these feature vectors. In our work, we use the following simple visual features:

$$F(x,y) = \left[x \quad y \quad L(x,y) \quad a(x,y) \quad b(x,y) \quad \left|\frac{\partial I(x,y)}{\partial x}\right| \quad \left|\frac{\partial I(x,y)}{\partial y}\right| \right]^T \tag{2}$$

where L,a,b denotes the color of the pixel in $L^*a^*b^*$ color space, $\left|\frac{\partial I}{\partial x}\right|, \left|\frac{\partial I}{\partial y}\right|$ encodes the edge orientations, and (x,y) denotes the pixel location. Hence, the covariance descriptor of

[1] We compute the SLIC superpixels by using its implementation in the VLFeat Library, http://www.vlfeat.org/.

a superpixel is computed as a 7 × 7 matrix. As illustrated in Fig. 3, superpixels with similar texture and local structures are described by similar covariance matrices.

Covariance matrices do not live on an Euclidean space which makes learning a visual dictionary from them very hard. Therefore, we use the idea offered by Hong *et al.* [17] to transform the covariance matrices into an Euclidean vector space. Mathematically speaking, let $\mathbf{C}$ be a $d \times d$ covariance matrix, a unique set of points $\mathcal{S} = \{\mathbf{s}_i\}$, referred to as Sigma Points, can be computed as:

$$\mathbf{s}_i = \begin{cases} \eta\sqrt{d}\mathbf{L}_i & \text{if } 1 \leq i \leq d \\ -\eta\sqrt{d}\mathbf{L}_i & \text{if } d+1 \leq i \leq 2d \end{cases} \tag{3}$$

with $\mathbf{L}_i$ denoting the ith column of the lower triangular matrix $\mathbf{L}$ obtained with the Cholesky decomposition $\mathbf{C} = \mathbf{L}\mathbf{L}^T$ and η being a scalar which is taken $\eta = \sqrt{2}$ as suggested in [17]. We then define the final descriptor of a superpixel by simply concatenating the mean vector of the features $\boldsymbol{\mu}$ and the elements of $\mathcal{S}$:

$$\mathbf{x}(\boldsymbol{\mu}, \mathbf{C}) = (\boldsymbol{\mu}, \mathbf{s}_1, \ldots, \mathbf{s}_d, \mathbf{s}_{d+1}, \ldots, \mathbf{s}_{2d})^T \tag{4}$$

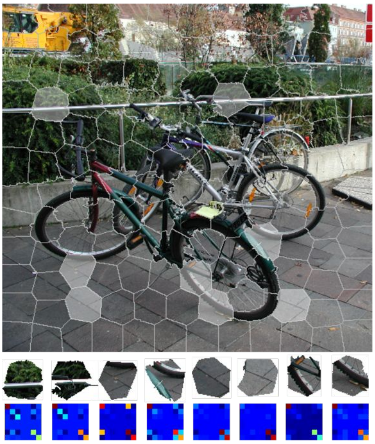

Figure 3: Superpixels having similar visual characteristics are represented by similar covariance descriptors.

3.2 CRF and Dictionary Learning for Saliency Estimation

We approach top-down saliency estimation as an image labeling problem in which a higher saliency score is assigned to superpixels corresponding to target objects. We construct a CRF model with nodes $\mathcal{V}$ representing the superpixels and edges $\mathcal{E}$ describing the connections among them. The saliency map is determined by finding the maximum posterior $P(\mathbf{Y}|\mathbf{X})$ of labels $\mathbf{Y} = \{y_i\}_{i=1}^n$ given the set of superpixels $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$:

$$\log P(\mathbf{Y}|\mathbf{X}, \mathbf{D}, \theta) = \sum_{i \in \mathcal{V}} \psi_i(y_i, \mathbf{x}_i; \mathbf{D}, \theta) + \sum_{i \in \mathcal{V}} \gamma_i(y_i, \mathbf{x}_i; \theta) + \sum_{(i,j) \in \mathcal{E}} \phi_{i,j}(y_i, y_j, \mathbf{x}_i, \mathbf{x}_j; \theta) - \log Z(\theta, \mathbf{D}) \tag{5}$$

where $y_i \in \{1, -1\}$ denotes the binary label of node $i \in \mathcal{V}$ indicating the presence or absence of the target object, ψ_i are the dictionary potentials, γ_i are the objectness potentials, $\phi_{i,j}$ are the edge potentials, θ are the parameters of the CRF model, and $Z(\theta, \mathbf{D})$ is the partition function. The model parameters $\theta = \{\mathbf{w}, \beta, \rho\}$ include the parameter of the dictionary potentials $\mathbf{w}$, the parameter of the objectness potentials β and the parameter of the edge potential ρ. The dictionary $\mathbf{D}$ used in ψ_i encodes the prior knowledge about the target object category.

Dictionary potential. The unary potentials ψ_i in our model depend on latent sparse variables defined over a trained discriminative dictionary $\mathbf{D}$. We use these sparse variables to learn a linear classifier, and use this classifier directly as our unary potential so that

$$\psi_i(y_i, \mathbf{x}_i; \mathbf{D}, \theta) = -y_i \mathbf{w}^T \boldsymbol{\alpha}_i \tag{6}$$

with $\boldsymbol{\alpha}_i$ denoting the sparse code of superpixel $\mathbf{x}_i$ which is computed using

$$\boldsymbol{\alpha}_i(\mathbf{x}_i, \mathbf{D}) = \arg\min_{\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}\|^2 + \lambda\|\boldsymbol{\alpha}\|_1 \tag{7}$$

where λ is a parameter which controls the sparsity.

Objectness potential. The objectness potential measures the likelihood that a superpixel belongs to an object in a class-independent manner. It is based on the objectness measure in [4] and obtained by sampling windows and accumulating their objectness scores. Since this procedure returns a pixelwise objectness map, we then take the average of objectness score of pixels inside each superpixel to obtain a superpixelwise map. Accordingly, we define the objectness potential $\gamma_i(y_i, \mathbf{x}_i; \theta)$ as:

$$\gamma_i(y_i, \mathbf{x}_i; \theta) = -\beta y_i \left(2P(obj|\mathbf{x}_i) - 1\right) \tag{8}$$

where $P(obj|\mathbf{x}_i)$ is the objectness score of superpixel $\mathbf{x}_i$, and β denotes the parameter of this potential function.

Edge potential. The pairwise edge potential $\phi_{i,j}$ models the interaction between two labels y_i, y_j of two neighboring superpixels. It has the form of the Potts model:

$$\phi_{i,j}(y_i, y_j, \mathbf{x}_i, \mathbf{x}_j; \theta) = \rho\left(1 - \delta(y_i - y_j)\right) \tag{9}$$

with δ denoting the unit impulse function. This potential gives a low energy value for equal labels and penalize any pair of different labels uniformly.

We simultaneously learn the CRF parameters θ and the dictionary $\mathbf{D}$ using a set of training images $\mathcal{X} = \{\mathbf{X}^{(m)}\}_{m=1}^{M}$ where the pixel-level ground truth annotations $\mathcal{Y} = \{\mathbf{y}^{(m)}\}_{m=1}^{M}$ are available. More specifically, we label a superpixel as positive (belonging to the object class) if the overlap ratio with the pixel-level ground truth annotation is larger than 95%, and as negative otherwise. Then, we learn $\mathbf{D}$ and θ by optimizing the following problem:

$$(\mathbf{D}^*, \theta^*) = \arg\max_{\mathbf{D},\theta} \prod_{m=1}^{M} P(\mathbf{Y}^{(m)}|\mathbf{X}^{(m)}, \mathbf{D}, \theta)\,. \tag{10}$$

using the iterative procedure described in [39] by alternating between dictionary and parameter updates. Jointly learning the CRF parameters and the dictionary provides a better discriminative power to separate the target object from the background.

We test the proposed model under three different settings. Firstly, we set the parameter of the objectness potential β to 0 and learn the CRF parameters $\mathbf{w}$ and ρ, and the superpixel-based dictionary $\mathbf{D}$ (setting 1). This can be seen as the superpixel-based extension of [39] as the objectness potential is excluded. Secondly, we learn all the CRF parameters θ and the dictionary $\mathbf{D}$ simultaneously by alternating between the update equations (setting 2). Lastly, we extend the first setting by determining the parameter of the objectness potential β later via cross-validation, while keeping the learned dictionary $\mathbf{D}$ and the other CRF parameters fixed (setting 3).

Once we have learnt the dictionary $\mathbf{D}$ and the optimal CRF parameters θ, saliency map of a test image can be computed according to Eq. (6) by first estimating the latent sparse variables $\alpha_{\mathbf{i}}$'s through Eq. (7) and then referring the labels by applying graph cuts. Our experiments show that the proposed approach has a superior performance to competing approaches (Section 4). We also observe that including the bottom-up generic objectness cues in our CRF model greatly improves the prediction quality.

4 Experiments

In Sections 4.1 and 4.2, we give an overview of our results on two benchmark datasets, namely Graz02 [26] and PASCAL VOC 2007 [15]. These datasets are both challenging since there contain large intra-class variations among object instances, and there are severe background clutter and occlusions.

4.1 Graz-02

Dataset and experimental setup. We conducted our first group of experiments on the Graz-02 dataset [26]. This dataset is composed of 1200 images from three different object categories (bike, car and people) and a background class. Each category has 300 images with additional pixel-level object annotations. For the experiments, we follow the standard experimental setup and use the odd numbered 150 images from each target object class and the odd numbered 150 images from the background class in the training, the even numbered images in the testing phase. We calculate superpixels by setting region size parameter to 0.05 and regularizer parameter to 0.1. For the training step we prefer to use same setup with [39]; 512 visual words and we set the initial learning rate to $1e$-3, the weight penalty to $1e$-5, the λ parameter, which is used in the sparse coding formulation, to 0.15. We train all the models with 20 iterations.

Evaluation and results. Table 1 reports the pixel-wise precision rates at equal error rates (EER) where precision is equal to recall for the proposed model, together with the results of the state-of-the-art bottom-up [23, 27, 38] and top-down [3, 20, 24, 39] saliency models. We use the aforementioned experimental setup for all methods. Incorporating top-down knowledge about the object of interest improves the overall pixel-level rates by a great extent. As expected, compared to the bottom-up saliency models [23, 27, 38] that do not utilize any knowledge about the appearance characteristics of object classes, the top-down saliency models have higher EER rates for all object categories.

In general, all of our models provide highly competitive results against the state-of-the-art top-down models. The results of our first setting clearly demonstrate the advantage of using superpixels rather than image patches as utilized in [39]. Jointly learning superpixel-based dictionaries and all the CRF parameters (setting 2), on the other hand, does not introduce any improvement in terms of EER. This may be due to the fact that inclusion of the objectness prior in the joint learning process decreases the discriminative power of the learned dictionary. However, determining the parameter of the objectness potential after learning the dictionary and the other CRF parameters (setting 3) further boosts the performance, and helps to achieve the best performance for all object categories, outperforming all other top-down saliency models.

Fig. 4 presents some qualitative results. We provide more experimental results in the supplementary material. From these results, we see that the bottom-up saliency models and the generic objectness model produce very poor saliency maps. While the bottom-up models

	Bike	Car	People
Margolin [23]	25.6	16.9	17.4
Perazzi *et al.* [27]	11.4	13.8	14.3
Yang and Zhang [38]	14.8	13.7	14.9
Objectness [4]	53.5	48.3	43.5
Aldavert *et al.* [3]	71.9	64.9	58.6
Khan and Tappen [20]	72.1	-	-
Marszalek and Schmid [24]	61.8	53.8	44.1
Yang and Yang [39]	62.4	60.0	62.0
Our approach (setting 1)	71.9	61.9	65.5
Our approach (setting 2)	71.7	62.0	64.9
Our approach (setting 3)	**73.9**	**68.4**	**68.2**

Table 1: EER results on the Graz-02 dataset.

select the image regions/objects having distinct characteristics from their surroundings, the generic objectness model has a tendency to pick all the objects exist in the images, and thus they can not locate the actual object of interests in a satisfactory way. The model of Yang and Yang [39] generates much better top-down saliency maps, however it does not capture the object boundaries well and selects some parts of the background as object as it carries out its analysis at patch level. Our third setting yields the most accurate saliency maps.

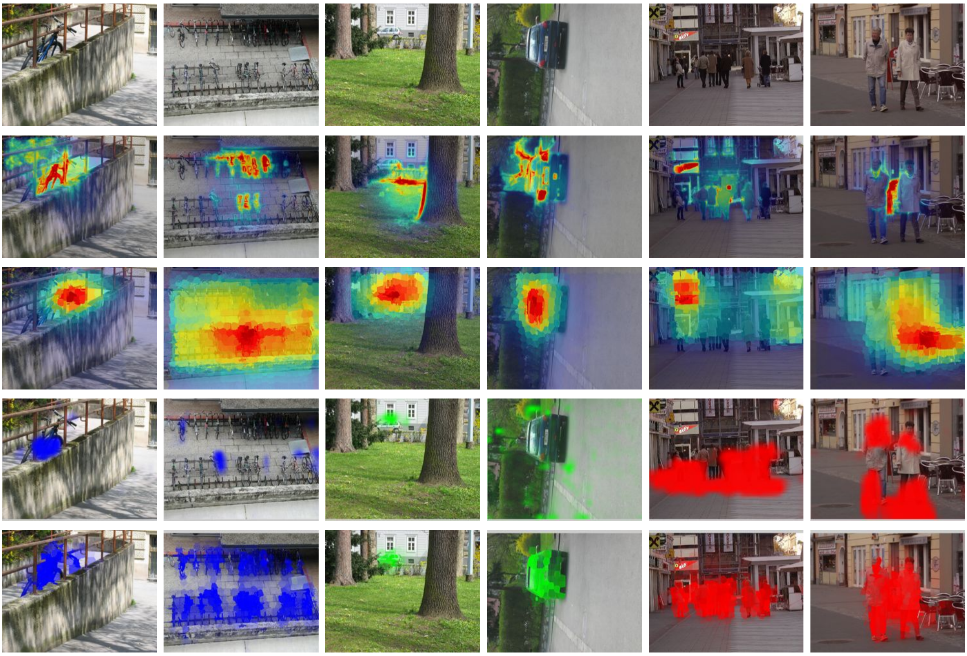

Figure 4: Sample results on the Graz-02 dataset. From top to bottom, input images and the saliency maps generated by the bottom-up model [23], the objectness model [4], the top-down model of [39] and the proposed approach. Bikes, cars and people are highlighted with blue, green and red, respectively for the top-down models.

4.2 PASCAL VOC 2007

Experimental setup. We used the PASCAL VOC 2007 dataset [15] in our second set of experiments. This dataset is more challenging than the Graz-02 dataset, with more number of visual object categories and more variations in the images. In the dataset, there are 9963 images from 20 different object categories and a background class. Most of these images, however, lack ground-truth pixel-level annotations. Only 632 of them have object level annotations, and this makes learning proper appearance models of the objects very hard. Another factor that makes the dataset difficult is that some object categories have parts with very similar appearances, e.g., wheels are shared across the bicycle, bus, car and motorbike categories. We conduct our experiments using the same training and test splits defined for the PASCAL VOC 2007 object segmentation challenge. From the pixelwise annotated 632 images, 422 of them are used for training and 210 of them are employed for testing. However, as also reported in [39], this results in very unbalanced training sets for most of the visual categories. The number of negative examples is generally much larger than that of positive examples. To work on a more balanced set and to obtain better generalizations, we generate extra object annotations from the remaining (pixelwise) unannotated sample images with the GrabCut [30] segmentation method using the existing bounding box annotations and include them to our training set. In this way, we can complete our training image number to 150 for each object class to balance training sets just in Graz-02 dataset. We only generate results for our third setting. We use the same set of parameters reported for the Graz-02 dataset except the region size parameter which is set to 0.07 to reduce computation time.

Evaluation and results. In Table 2, we present results obtained for the proposed model (setting 3) and the model suggested in [39]. Here, we again report category-level EER rates for all the 210 annotated test images. Compared to [39], our saliency maps are much better for all categories. The only exception is the person category for which our approach performs slightly worse than [39]. It should also be noted that the EER rated reported in [39] are on the patch level whereas our rates are given for the pixelwise saliency maps. We observe that our method does not perform well for some classes such as sofa, bottle, tv-monitor, chair because of the lack of distinctive characteristics in these classes. We present some representative qualitative results in Fig. 5

5 Conclusion

In this paper, we have presented a novel approach for generating class-specific top-down saliency maps for real-world images. The key point of the proposed method is to carry out the saliency estimation process over a set of superpixels, instead of at the pixel or patch level. We cast the problem as a graph labeling problem and jointly learn a CRF and a discriminative dictionary using superpixels. Our experimental evaluation on the Graz-02 and PASCAL VOC 2007 datasets shows that the proposed approach provide more accurate saliency maps as compared to the relevant previous work for various object categories.

Acknowledgments

This research was supported in part by The Scientific and Technological Research Council of Turkey (TUBITAK), Career Development Award 112E146.

	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow
Yang and Yang [39]	15.2	39.0	9.4	5.7	3.4	22.0	30.5	15.8	5.7	8
Our result	**49.4**	**46.6**	**33.7**	**60.9**	**26.1**	**51.8**	**35.1**	**64.9**	**21.1**	**34.8**
	dining table	dog	horse	motorbike	person	potted plant	sheep	sofa	train	tv-monitor
Yang and Yang [39]	11.1	12.8	10.9	23.7	**42.0**	2.0	20.2	10.4	24.7	10.5
Our result	**43.7**	**35.1**	**41.4**	**71.4**	32.6	**42**	**42.5**	**13.8**	**63.8**	**27.8**

Table 2: EER results on the PASCAL VOC 2007 dataset.

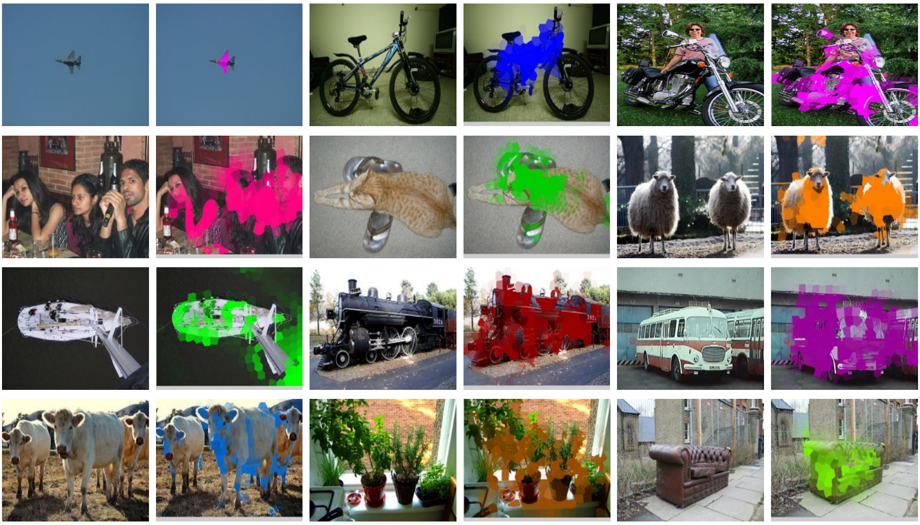

Figure 5: Sample results of our approach (setting 3) on the PASCAL VOC 2007 dataset.

References

[1] R. Achanta and S. Susstrunk. Saliency detection for content-aware image resizing. In *ICIP*, pages 1005–1008, 2009.

[2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(11):2274–2282, 2012.

[3] D. Aldavert, A. Ramisa, R.L. de Mantaras, and R. Toledo. Fast and robust object segmentation with the integral linear classifier. In *CVPR*, pages 1046–1053, 2010.

[4] B. Alexe, T. Deselares, and V. Ferrari. What is an object? In *CVPR*, 2010.

[5] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):185–207, 2013.

[6] A. Borji, D. N. Sihite, and L. Itti. Salient object detection: A benchmark. In *ECCV*, pages 414–429, 2012.

[7] N.J. Butko, Lingyun Zhang, G.W. Cottrell, and Javier R. Movellan. Visual saliency model for robot cameras. In *ICRA*, pages 2398–2403, 2008.

[8] J. Carreira and C. Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In *CVPR*, 2010.

[9] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr. BING: Binarized normed gradients for objectness estimation at 300fps. In *CVPR*, 2014.

[10] R. G. Cinbis, J. Verbeek, and C. Schmid. Segmentation driven object detection with fisher vectors. In *ICCV*, 2013.

[11] M. Donoser, M. Urschler, M. Hirzer, and H. Bischof. Saliency driven total variation segmentation. In *ICCV*, pages 817–824, 2009.

[12] K. A. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva. Modeling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition*, 17 (6-7):945–978, 2009.

[13] I. Endres and D. Hoiem. Category independent object proposals. In *ECCV*, 2010.

[14] E. Erdem and A. Erdem. Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of Vision*, 13(4):1–20, 2013.

[15] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[16] B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *ICCV*, pages 670–677, 2009.

[17] X. Hong, H. Chang, S. Shan, X. Chen, and W. Gao. Sigma set: A small second order statistical region descriptor. In *CVPR*, pages 1802–1809, 2009.

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

[19] C. Kanan, M. H. Tong, L. Zhang, and G. W. Cottrell. SUN: Top-down saliency using natural statistics. *Visual Cognition*, 17(6-7):979–1003, 2009.

[20] N. Khan and M.F. Tappen. Discriminative dictionary learning with spatial priors. In *ICIP*, pages 166–170, 2013.

[21] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang. A generic framework of user attention model and its application in video summarization. *IEEE Trans. Mult.*, 7(5):907–919, 2005.

[22] S. Manen, M. Guillaumin, and L. V. Gool. Prime object proposals with randomized Prim's algorithm. In *ICCV*, 2013.

[23] R. Margolin, A. Tal, and Zelnik-Manori L. What makes a patch distinct? In *CVPR*, 2009.

[24] M. Marszalek and C. Schmid. Accurate object recognition with shape masks. *Int. J. Comput. Vision*, 97(2):191–209, 2012.

[25] V. Navalpakkam and L. Itti. An integrated model of top-down and bottom-up attention for optimizing detection speed. In *CVPR*, volume 2, pages 2049–2056, 2006.

[26] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer. Generic object recognition with boosting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(3):416–431, 2006.

[27] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, pages 733–740, 2012.

[28] E. Rahtu, J. Kannala, and M. Blaschko. Learning a category independent object detection cascade. In *ICCV*, 2011.

[29] P. Rantalankila, J. Kannala, and E. Rahtu. Generating object segmentation proposals using global and local search. In *CVPR*, 2014.

[30] C. Rother, V. Kolmogorov, and A. Blake. GrabCut: Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, 2004.

[31] U. Rutishauser, D. Walther, C. Koch, and P. Perona. Is bottom-up attention useful for object recognition? In *CVPR*, volume 2, pages II–37–II–44 Vol.2, June 2004.

[32] G. Sharma, F. Jurie, and C. Schmid. Discriminative spatial saliency for image classification. In *CVPR*, pages 733–740, 2012.

[33] C. Siagian and L. Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(2):300–312, 2007.

[34] A. Torralba, A Oliva, M.S. Castelhano, and J.M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, 113(4):766–786, 2006.

[35] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. *ECCV*, pages 589–600, 2006.

[36] K. E. van de Sande, J. R. Uijlings, T. Gevers, and A. W. Smeulders. Segmentation as selective search for object recognition. In *ICCV*, 2011.

[37] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Netw.*, 19 (9):1395–1407, 2006.

[38] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, pages 3166–3173, 2013.

[39] J. Yang and M.-H. Yang. Top-down visual saliency via joint CRF and dictionary learning. In *CVPR*, pages 2296–2303, 2012.