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

We present a method for learning top-down visual saliency, which is well-suited to locate objects of interest in complex scenes. Our approach is inspired in part by the recent dictionary-based top-down saliency approaches [4, 9] and the new superpixel-based bottom-up salient object detection methods [5, 7, 8]. Specifically, we approach top-down saliency estimation as an image labeling problem in which higher saliency scores are assigned to the image locations corresponding to the target object.

Given a set of training images containing object level annotations, we first segment the images into superpixels. Additionally, we extract 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 the corresponding dictionaries learned from data, estimate the objectness map and use the CRF model to infer saliency scores (see Figure 1).

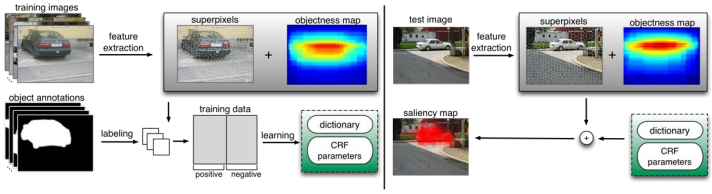


Figure 1: System overview.

Superpixel representation. We segment the images into superpixels and represent them by means of the the first and the second order statistics of simple visual features including color, edge orientation and spatial information. For this step, we employ the sigma points descriptor [3] which provides a compact and effective way of encoding statistical relationships among simple visual features.

CRF and dictionary learning for saliency estimation. 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 $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$:

$$\log P(\mathbf{Y}|\mathbf{X},\mathbf{D},\theta) = \sum_{i\in\mathcal{V}} \psi_i(y_i,\mathbf{x}_i;\mathbf{D},\theta) + \sum_{i\in\mathcal{V}} \gamma_i(y_i,\mathbf{x}_i;\theta) + \sum_{(i,j)\in\mathcal{E}} \phi_{i,j}(y_i,y_j,\mathbf{x}_i,\mathbf{x}_j;\theta) - \log Z(\theta,\mathbf{D}) \quad (1)$$

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.

We test the proposed model under three different settings. In setting 1, we ignore objectness potential and learn discriminative dictionaries and CRF model at superpixel level. In setting 2, we jointly learn dictionary and CRF model by including objectness prior. Setting 3 is extended version of the first one which determines the parameter of the objectness potential β later via cross-validation, while keeping the learned dictionary D and the other CRF parameters fixed.

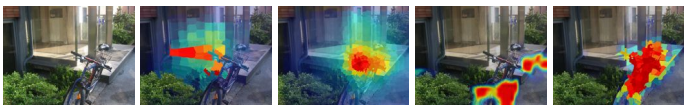
We demonstrate the effectiveness of our approach by comparing it with several bottom-up and top-down models and a generic objectness approach (see Table 1 and 2 for overall results and Figure 2 for a sample comparison). In general, bottom-up models and generic objectness approach do not capture the object of interest due to lack of prior knowledge about the object of interest, and the patch-based top-down saliency models either partly capture the target objects or provide very coarse localizations of the target objects. Our saliency model results in considerably better top-down saliency maps.

	Bike	Car	People
Margolin [5]	25.6	16.9	17.4
Perazzi [7]	11.4	13.8	14.3
Yang and Zhang [8]	14.8	13.7	14.9
Objectness [2]	53.5	48.3	43.5
Aldavert [1]	71.9	64.9	58.6
Khan and Tappen [4]	72.1	-	-
Marszalek and Schmid [6]	61.8	53.8	44.1
Yang and Yang [9]	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.

	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow
Yang and Yang [9]	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 [9]	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.



(a) (b) (c) (d) (e)

Figure 2: Results for the *look for a bike* task. (a) Input image, and the results of (b) a bottom-up saliency model [8], (c) the objectness map generated by [2], (d) the top-down saliency model of [9] and (e) our approach.

[1] 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.

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

[3] 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.

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

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

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

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

[8] 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.

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