Top-down Saliency with Locality-constrained Contextual Sparse Coding

Hisham Cholakkal
hisham002@ntu.edu.sg
Deepu Rajan
http://www3.ntu.edu.sg/home/ASDRajan
Jubin Johnson
jubin001@ntu.edu.sg

School of Computer Engineering
Nanyang Technological University
Singapore

We propose a locality-constrained contextual sparse coding (LCCSC) for top-down saliency estimation where higher saliency scores are assigned to the image locations corresponding to the target object. Three locality constraints are integrated in to this novel sparse coding. First is the spatial or contextual locality constraint in which features from adjacent regions have similar code, second is the feature-domain locality constraint in which similar features have similar code, and third is the category-domain locality constraint in which features are coded using similar atoms from each partition of the dictionary, where each partition corresponds to an object category.

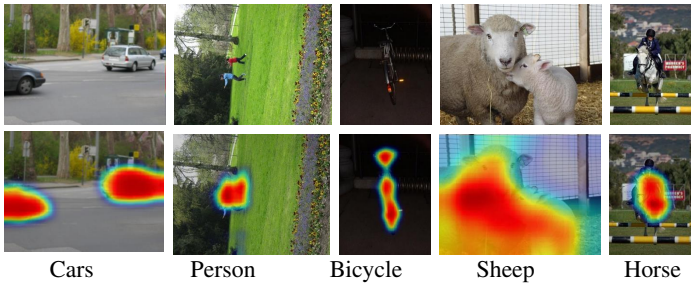

Figure 1: Our Saliency maps on Graz-02 and PASCAL VOC-07 datasets.

The pipeline of the proposed method is shown in fig. 2, which is similar to the widely used sparse coded spatial pyramid matching (ScSPM) image classifier, i.e., feature extraction, feature coding, pooling and feature classifier. Spatial neighborhood of a feature is divided into a regular grid and the codes in each cell of the grid are max-pooled individually. These max-pooled vectors are vertically concatenated to form a context max-pooled vector representing the patch. Logistic regression-based feature classifier is learnt using these context max-pooled vectors. For saliency inference on a test image, the class-conditional probability of context max-pooled vectors are estimated from the learnt logistic regression model. This probability is the saliency value of a patch in the image. The pixel-level saliency map is obtained using a novel Gaussian-weighted interpolation of the patch-level saliency map.

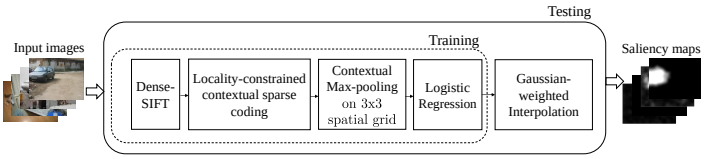

Figure 2: Training and testing of the proposed top-down saliency.

Locality-constrained contextual sparse coding (LCCSC) Various coding schemes aim for specific objectives like sparsity, feature-domain locality and spatial-domain locality. Here, all these desired properties are integrated into a single objective function. LCCSC ensures that features representative of salient regions are not ignored even if they are not discriminative.

Given a feature vector f and dictionary D, LCCSC coding searches for the codeword z that satisfies the following criteria:

$$\arg\min_{z} \|f - Dz\|_2 + \lambda_1 \|z\|_1 + \lambda_2 \|z \odot h_\pi\|_2 + \lambda_3 \sum_{j=1}^{c} (|\frac{\|z\|_0}{c} - \|z \odot col_j[\rho]\|_0|); \quad (1)$$

The first two terms are the conventional sparse coding of feature f with l_1 constraint. The third term imposes locality constraint in the feature domain as well as in the spatial domain. The fourth term imposes category-domain locality, i.e the number of atoms that contribute to the non-zero values of z are distributed among all object categories. Detailed discussion about eq. (1) and an approximate solution is available in the paper.

Table 1: Pixel-level recision rates (%) at EER on Graz-02.

	Bicycle	Car	Person	*Mean*
Objectness	53.5	48.3	43.5	48.43
Aldavert *et al.*	71.9	64.9	58.6	65.13
Khan and Tappen	72.1	-	-	-
Marszalek and Schmid	61.8	53.8	44.1	53.23
Yang and Yang [2]	62.1	60.0	62.0	61.46
Kocak *et al.* [1]	73.9	68.4	**68.2**	70.16
LCCSC-pooled (*upsampling of [2]*)	73.41	68.6	61.25	67.75
LCCSC-pooled (*Gaussian-weighted interpolation*)	**76.19**	**71.2**	64.13	**70.49**

We test our method on two challenging datasets - Graz-02 and PASCAL VOC-07. Table 1 compares the pixel-level results of our method with other top-down saliency models. The proposed method is called as LCCSC-pooled, where pooled indicates contextual max pooling of the code. It is to be noted that we achieve state-of-the-art performance by using simpler feature coding and contextual max-pooling in comparison with computationally complex dictionary learning and graph-based approaches of [1, 2].

As in [2], all models are evaluated on the entire 210 segmentation test images of PASCAL VOC-07 dataset, irrespective of the presence or absence of target. We outperform [2] in 15 out of 20 classes. On averaging across all classes, in patch-level, we achieve a mean precision rate at EER of 23.4% which is better than 16.15% of [2].

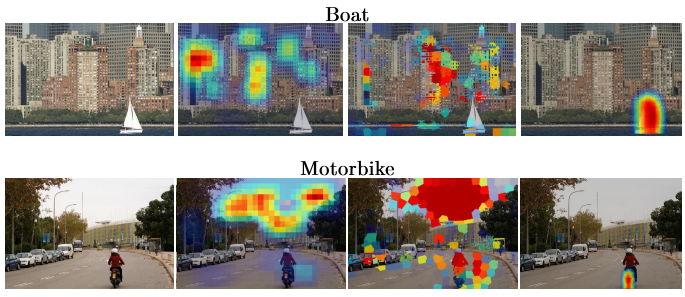

Figure 3: Comparison with state-of-the-art top-down saliency approaches on challenging test images. (a) input image, (b) Saliency maps of Yang and Yang [2], (c) Kocak *et al.* [1] and (d) the proposed method.

[1] A. Kocak, K. Cizmeciler, A. Erdem, and E. E. Top down saliency estimation via superpixel-based discriminative dictionaries. In *BMVC*, 2014.

[2] J. Yang and M.-H. Yang. Top-down visual saliency via joint crf and dictionary learning. In *CVPR*, 2012.