Salient Object Detection via Bootstrap Learning

Na Tong[1], Huchuan Lu[1], Xiang Ruan[2] and Ming-Hsuan Yang[3]
[1]Dalian University of Technology. [2]OMRON Corporation. [3]University of California at Merced.

Salient object detection methods can be categorized as bottom-up stimuli-driven [2, 6] and top-down task-driven [4, 8] approaches. In this paper, we propose a novel algorithm via bootstrap learning [5] in which both weak and strong models are exploited. To address the problems of noisy detection results and limited representations from bottom-up methods, we present a learning approach to exploit multiple features. However, unlike existing top-down learning-based methods, the proposed algorithm is bootstrapped with samples from a bottom-up model, thereby alleviating the time-consuming off-line training process or labeling positive samples manually. Figure 1 shows the main steps of the proposed salient object detection algorithm.

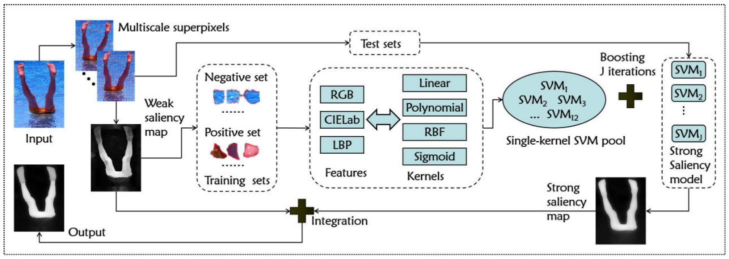

Figure 1: Bootstrap learning for salient object detection.

First, we construct a weak saliency model by exploiting the contrast between each region and the regions along the image border based on three descriptors including the RGB, CIELab and Local Binary Pattern (LBP) features. In addition, the center-bias and dark channel priors are further exploited to better estimate saliency maps. An input image is segmented into N superpixels, $\{c_i\}, i = 1, \ldots, N$. The regions along the image border are represented as $\{n_j\}, j = 1, \ldots, N_B$. The coarse saliency value for the region c_i is constructed by

$$f_0(c_i) = g(c_i) \times S_d(c_i) \times \sum_{\kappa \in \{F_1, F_2, F_3\}} \left(\frac{1}{N_B} \sum_{j=1}^{N_B} d_\kappa(c_i, n_j) \right), \tag{1}$$

where $d_\kappa(c_i, n_j)$ is the Euclidean distance between region c_i and n_j in the feature space that κ represents, i.e., the RGB (F_1), CIELab (F_2) and LBP (F_3) texture features respectively. In addition, $S_d(c_i)$ and $g(c_i)$ denote the dark channel prior and center prior for the region c_i. We use a simple yet effective algorithm based on the Graph Cut method [1], to construct the continuous and smoothed weak saliency map, from which the training set for the strong classifier is selected.

Then, we present a method similar to the Multiple Kernel Boosting (MKB) [7] method to include multiple kernels of different features. We treat SVMs with different kernels as weak classifiers and then learn a strong classifier using the boosting method. Note that we restrict the learning process to each input image to avoid the heavy computational load of extracting features and learning kernels for a large amount of training data (as required in several discriminative methods [4] in the literature for saliency detection).

For each image, we have the training samples $\{r_i, l_i\}_{i=1}^{H}$, where r_i is the i-th sample, l_i represents the binary label of the sample and H indicates the number of the samples. In this paper we use the boosting algorithm instead of the simple combination of single-kernel SVMs in the MKL method:

$$Y(r) = \sum_{j=1}^{J} \beta_j z_j(r). \tag{2}$$

In order to compute the parameters β_j, we use the Adaboost method and the parameter J in (2) denotes the number of iterations of the boosting process. We consider each SVM as a weak classifier and the final strong classifier $Y(r)$ is the weighted combination of all the weak classifiers. Starting with uniform weights, $\omega_1(i) = 1/H, i = 1, 2, \ldots, H$, for the SVM classifiers, we obtain a set of decision functions $\{z_m(r)\}, m = 1, 2, \ldots, M$. At the j-th iteration, we compute the classification error for each of the weak classifiers,

$$\varepsilon_m = \frac{\sum_{i=1}^{H} \omega(i) |z_m(r_i)| (\mathrm{sgn}(-l_i z_m(r_i)) + 1)/2}{\sum_{i=1}^{H} \omega(i) |z_m(r_i)|}, \tag{3}$$

where $\mathrm{sgn}(x)$ is the sign function. We locate the decision function $z_j(r)$ with the minimum error ε_j, i.e., $\varepsilon_j = \min_{1 \le m \le M} \varepsilon_m$. Then the combination coefficient β_j is computed by $\beta_j = \frac{1}{2} \log \frac{1-\varepsilon_j}{\varepsilon_j} \cdot \frac{1}{2}(\mathrm{sgn}(\log \frac{1-\varepsilon_j}{\varepsilon_j}) + 1)$. Note that β_j must be larger than 0, indicating $\varepsilon_j < 0.5$, which accords with the basic hypothesis that the boosting method could make the weak classifiers into a strong one. In addition, we update the weight using the following equation,

$$\omega_{j+1}(i) = \frac{\omega_j(i) e^{-\beta_j l_i z_j(r_i)}}{2\sqrt{\varepsilon_j(\varepsilon_j - 1)}}. \tag{4}$$

After J iterations, all the β_j and $z_j(r)$ are computed and we have a boosted classifier (2) as the saliency model learned directly from an input image. We apply this strong saliency model to the test samples (based on all the superpixels of an input image), and a pixel-wise saliency map is thus generated.

To improve the accuracy of the map, we first use the Graph Cut method to smooth the saliency detection results. Next, we obtain the strong saliency map by further enhancing the saliency map with the guided filter [3] as it has been shown to perform well as an edge-preserving smoothing operator.

The accuracy of the saliency map is sensitive to the number of superpixels as salient objects are likely to appear at different scales. To deal with the scale problem, we generate four layers of superpixels with different granularities.

The weak map is likely to detect fine details and to capture local structural information due to the contrast-based measure. In contrast, the strong map works well by focusing on global shapes for most images except the case when the test background samples have similarity with the positive training set or large differences compared to the negative training set, or vice versa for the test foreground sample. Thus we integrate the weak and strong maps by a weighted combination to incorporate the complementary properties of these two maps.

Extensive experiments on six benchmark datasets demonstrate that the proposed bootstrap learning algorithm performs favorably against the state-of-the-art saliency detection methods. Furthermore, we show that the proposed bootstrap learning approach can be easily applied to other bottom-up saliency models for significant improvement.

[1] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 23(11):1222–1239, 2001.

[2] Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi-Min Hu. Global contrast based salient region detection. *PAMI*, 37(3):569–582, 2015.

[3] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. In *ECCV*, 2010.

[4] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. Salient object detection: A discriminative regional feature integration approach. In *CVPR*, 2013.

[5] Benjamin Kuipers and Patrick Beeson. Bootstrap learning for place recognition. In *AAAI*, 2002.

[6] Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via dense and sparse reconstruction. In *ICCV*, 2013.

[7] Fan Yang, Huchuan Lu, and Yen-Wei Chen. Human tracking by multiple kernel boosting with locality affinity constraints. In *ACCV*, 2010.

[8] Jimei Yang and Ming-Hsuan Yang. Top-down visual saliency via joint CRF and dictionary learning. In *CVPR*, 2012.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.