Visual Saliency Based on Multiscale Deep Features

Guanbin Li, Yizhou Yu
Department of Computer Science, The University of Hong Kong

Visual saliency attempts to determine the amount of attention steered towards various regions in an image by the human visual and cognitive systems. It is thus a fundamental problem in psychology, neural science, and computer vision. Visual saliency has been incorporated in a variety of computer vision and image processing tasks to improve their performance. Such tasks include image cropping, retargeting, and summarization.

In this paper, we discover that a high-quality visual saliency model can be learned from multiscale features extracted using deep convolutional neural networks (CNNs), which have had many successes in visual recognition tasks. For learning such saliency models, we introduce a neural network architecture, which has fully connected layers on top of CNNs responsible for extracting features at three different scales. We then propose a refinement method to enhance the spatial coherence of our saliency results. Finally, aggregating multiple saliency maps computed for different levels of image segmentation can further boost the performance, yielding saliency maps better than those generated from a single segmentation.

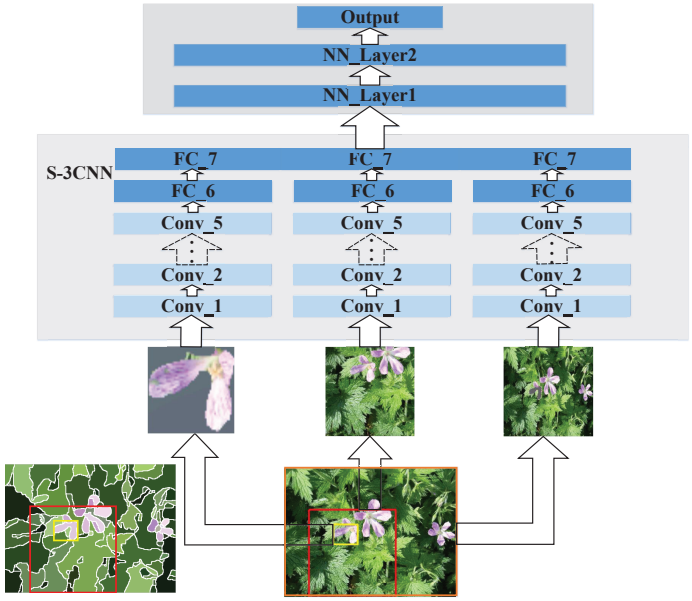

Figure 1: The architecture of our deep feature based visual saliency model.

As shown in Fig. 1, the architecture of our deep feature based model for visual saliency consists of one output layer and two fully connected hidden layers on top of three deep convolutional neural networks. Our saliency model requires an input image to be decomposed into a set of nonoverlapping regions. By definition, saliency is resulted from visual contrast as it intuitively characterizes certain parts of an image that appear to stand out relative to their neighboring regions or the rest of the image. Thus, to compute the saliency of an image region, our model evaluates the contrast between the considered region and its surrounding area as well as the rest of the image. Therefore, the three CNNs extract multiscale features for every image region from three nested and increasingly larger rectangular windows, which respectively encloses the considered region, its immediate neighboring regions, and the entire image. The features extracted from the three CNNs are fed into the two fully connected layers, each of which has 300 neurons. These fully connected layers play the role of a regressor that is capable of inferring the saliency score of every image region from the multiscale CNN features. The output of the second fully-connected layer

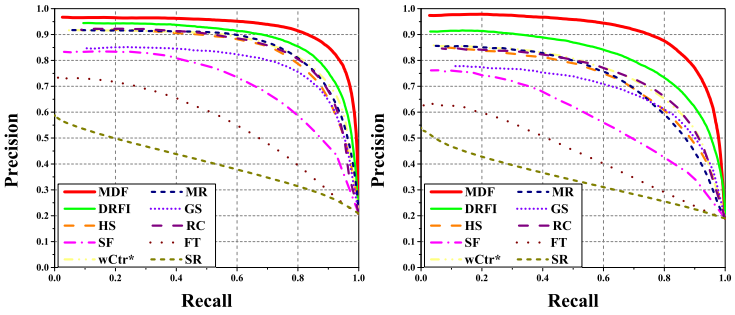

Figure 2: Precision-recall curves of 10 different methods, including ours (MDF), on two datasets: (Left) the MSRA-B dataset, and (Right) our new dataset (HKU-IS). Our method consistently outperforms other methods, including spectral residual (SR), frequency-tuned saliency (FT), saliency filters (SF), geodesic saliency (GS), hierarchical saliency, regional based contrast (RC [1]), manifold ranking (MR [4]), optimized weighted contrast (wCtr* [5]) and discriminative regional feature integration (DRFI [2]).

is fed into the output layer, which performs two-way softmax that produces a distribution over binary saliency labels. When generating a saliency map for an input image, we run our trained saliency model repeatedly over every region of the image to produce a single saliency score for that region. This saliency score is further transferred to all pixels within that region.

At present, the only large dataset that can be used for training a deep neural network was derived from the MSRA-B dataset [3]. This dataset has become less challenging over the years because images there typically include a single salient object located away from the image boundary. To facilitate research and evaluation of advanced saliency models, we have created a large dataset where an image likely contains multiple salient objects, which have a more general spatial distribution in the image. Our new saliency dataset, called HKU-IS, contains 4447 images with high-quality pixelwise annotations. In summary, 50.34% images in HKU-IS have multiple disconnected salient objects while this number is only 6.24% for the MSRA dataset; 21% images in HKU-IS have salient objects touching the image boundary while this number is 13% for the MSRA dataset; and the mean color contrast of HKU-IS is 0.69 while that of the MSRA dataset is 0.78.

Experimental results demonstrate that our proposed method (MDF) is capable of achieving state-of-the-art performance on all public benchmarks, improving the F-Measure by 5.0% and 13.2% respectively on the MSRA-B dataset and our new dataset (HKU-IS), and lowering the mean absolute error by 5.7% and 35.1% respectively on these two datasets. A quantitative comparison of precision-recall curves of 10 different methods is shown in Fig. 2, where our method consistently outperforms other methods.

[1] Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi-Min Hu. Global contrast based salient region detection. *TPAMI*, 2014.

[2] 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.

[3] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *TPAMI*, 33(2):353–367, 2011.

[4] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013.

[5] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. Saliency optimization from robust background detection. In *CVPR*, 2014.

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