Looking Beyond the Image: Unsupervised Learning for Object Saliency and Detection report

Paper's Authors

Parthipan Siva, Chris Russell, Tao Xiang, and Lourdes Agapito

Advisor

Dr. Maryam Abedi

Student

Mohammad Shahpouri

October
2022

Contents

List of Figures

List of Equations

1 Sampling-based saliency

To find a saliency map S_I for image I salient patches are defined as those have the least probability of being sampled from a set of images $\mathcal{D}_I$ similar to I. Here $\mathcal{D}_I$ includes the current image I and other images obtained from the corpus of unlabelled images $\mathcal{D}$ and patches are $n \times n$ regions around each image pixel. p_x a proportional number to the probablity of sampling patch x from $\mathcal{D}_I$.

A patch x from an image I is uniformly selected, then perturbing it by some noise in an informative feature space.

$$\begin{aligned} p_x &\propto \Pr(X = x|\mathcal{D}_I) \\ &= \int_{\mathcal{D}_I} \Pr(X = x|J)\ \mathrm{d}J \\ &= \int_{\mathcal{D}_I} \int_J \Pr(X = x|y)\ \Pr(y|J)\ \mathrm{d}y\ \mathrm{d}J \\ &= \int_{\mathcal{D}_I} \int_J \Pr(X = x|y)\ \mathrm{d}y\ \mathrm{d}J \end{aligned} \tag{1}$$

The noise is uniform and Gaussian:

$$p_x \propto \int_{\mathcal{D}_I} \int_J \exp\left(-\frac{d(x,y)^2}{\sigma^2}\right)\ \mathrm{d}y\ \mathrm{d}J \tag{2}$$

Here $d(x, y)$ is the Euclidean distance between the feature representations of patches x and y.

Then p_x can be approximated as:

$$p_x \approx \sum_{y \in N_m(x,\mathcal{D}_I\ \ \{I\})} \exp\left(\frac{-d(x,y)^2}{\sigma^2}\right) + \sum_{y \in N_m(x,\{I\})} \left(\frac{-d_I(x,y)}{\sigma^2}\right) \tag{3}$$

where $N_m(x, \mathcal{D}_I\backslash\ \{I\})$ are the m approximate nearest neighbors (ANNs) of patch x taken from all images, $\sigma = 1$ and $d_i(x, y)$ is:

$$d_I(x,y) = \left(\frac{d(x,y)^2}{1 + c\ \ \cdot (l(x) - l(y))^2}\right) \tag{4}$$

where $c = 3$ is a constant, $l()$ is the location of patches in normalised image coordinates.

Now p_x is approximated. A high value of p_x means that the patch x is common in the image corpus, and the saliency of patch x is obtained as:

$$S_x = 1 - p_x \tag{5}$$

where p_x , over all patches x in the image I, was normalised to the range $[0, 1]$. The saliency S_x are calculated at four different image scales $[1, .8, .5, .3,]$ and average the result over the four scales $\bar{S}_x$ as the patch saliency.

Post-Processing Two post-processing steps are applied to $\bar{S}_x$. First, as in [1], To encode immediate context information, high salient pixel locations $\mathcal{F} = \bar{S}_x > T$ are found and the saliency value at all pixel location i is weighted by their distance to $\mathcal{F}$. Second, a segmentation technique of [8] is applied to the saliency map to recover image boundary information. For each segment region, the average saliency from S_c is obtained and used as the final saliency value for that segment, producing the saliency map S_I.

$$S_c(i) = \bar{S}_x(i) \left(\sum_{y \in N_{64}(i,\mathcal{F})} \exp\left(-\frac{(l(i) - l(y))^2}{\sigma_l} \right) \right) \tag{6}$$

where $N_{64}(i, \mathcal{F})$ are the 64 nearest neighbours of i in $\mathcal{F}$, $l()$ is the normalised image coordinate of pixels. And $\sigma_l = 0.2$

Similar Images In Equation 3, approach of [9] is followed to select 20 similar images from $\mathcal{D}$ using Euclidean distance on GIST [10] descriptors and a 30×20 thumbnail image in Lab colour space.

Patch Features Two features are extracted from Each $n \times n$ patch. Lab color space of each patch of length $3n^2$, and 128 bin SIFT descriptor [11] of each $n \times n$ patch are calculated. Concatenation of the vectors result $3n^2 + 128$ feature descriptor of the patch. Lab feature patch size is 7×7 and a 4×4 block is used as SIFT descriptor.

2 Bounding box sampling

They propose a coherent sampling derives from non-maximum suppression (NMS) [5]. A set of all boxes $\mathcal{B}$ is considered, and seek $b_* \in \mathcal{B}$, the box that best explains the saliency of all bounding boxes in $\mathcal{B}$.

To find such a box, the boxes in $\mathcal{B}$ are drawn using a saliency weighted average BoW SIFT histogram:

$$\mu^{SIFT} = \frac{1}{\sum_{i=0}^{N} d_i} \sum_{i=0}^{N} d_i f^{SIFT}(b_i) \tag{7}$$

where $f^{SIFT}(b_i)$ is the dense SIFT BoW histogram representation of b_i and d_i is the saliency score of box b_i. Then to maximise the overlap with the salient boxes in $\mathcal{B}$ that will be suppressed, b_* is chosen as the box with the closest histogram to μ^{SIFT}.

$$b_* = \arg\min_{b_i} \|f^{SIFT}(b_i) - \mu^{SIFT}\|_2 \tag{8}$$

The saliency score d i for box b i is defined as:

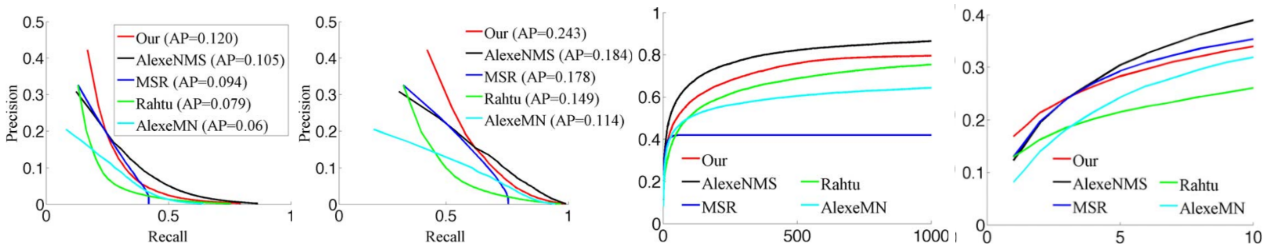
$$d_i = \frac{1}{|b_i|^r} \sum_{p \in b_i} S(p) - \frac{1}{|u_i|^r} \sum_{p \in u_i} S(p) \tag{9}$$

$|\cdot|$ refers to the size of the box in pixels, u_i is a buffer around the box b_i that ensures to select local maxima. r is a soft bias on the box size. When $r = 0$ the highest density box fills the image and if $r = 1$ the highest density box is typically only a single pixel wide. To sample boxes at different scales, instead of alternating between 4 explicit choices of scale [6], we alternate between sampling with a soft bias towards large scales with $r = 0.5$ and a bias towards smaller patches with $r = 0.75$.

3 Experimental results

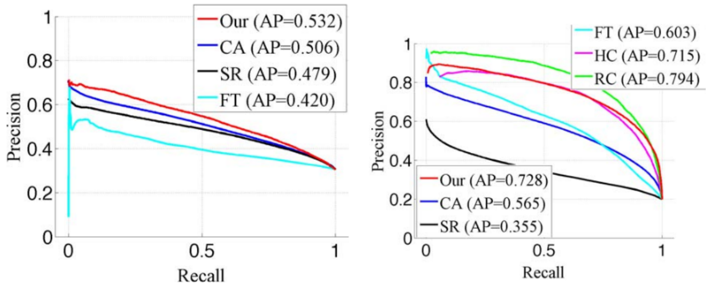
Their dataset consists of 98,000 images from LABLELME [12], PASCAL VOC 2007, AND 2009 [13].

3.1 Quantitative results



(a) All Objects (b) One Object per Image (c) Hit Rate (d) Hit Rate

Figure 1. Precision recall curve and recall vs number of object location proposal on the P ASCAL 2007 TrainVal dataset. (d) is a zoomed in view of (c). Best viewed in colour.



(a) VOC07 (b) MSRA

Figure 2. Per-pixel accuracy vs CA [1], SR [2], FT [3], HC [4], RC [4].

3.2 Qulitative results

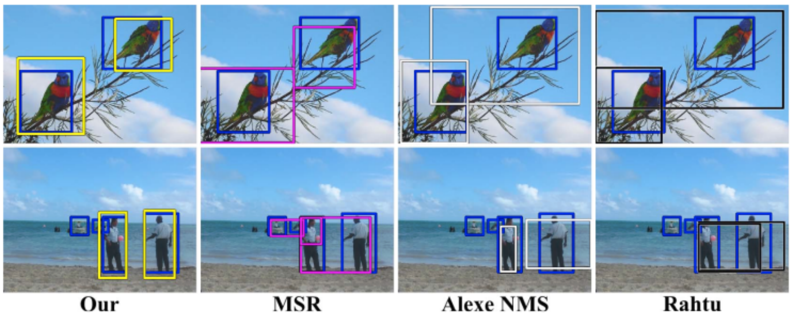


Figure 3. Best bounding boxes taken from the top 10 proposed object locations by our coherent sampling method (Our), MSR [5], Alexe et al. NMS [6], and Rahtu et al. [7]. Blue is ground truth.

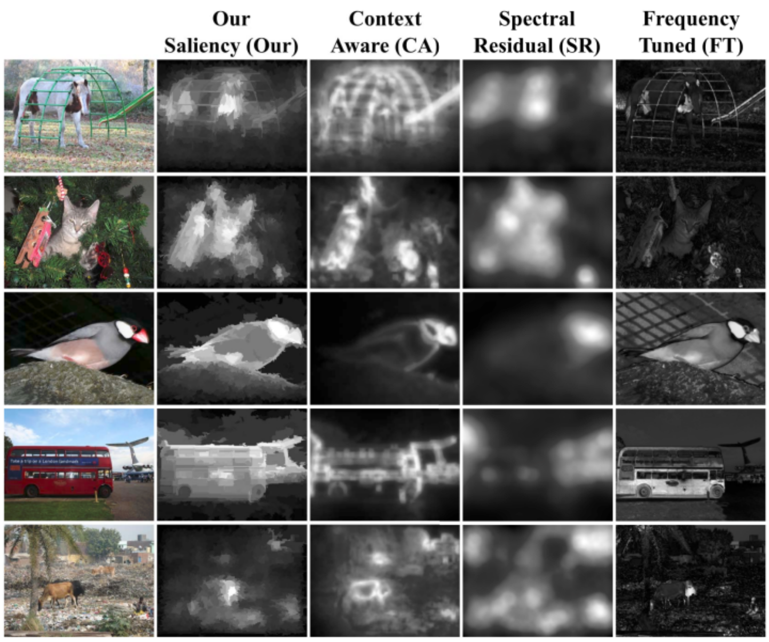


Figure 4. Our image saliency in comparison to CA [1], SR [2], and FT [3] methods.

References

[1] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1915–1926, 2012. ii, 2, 4, 5

[2] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007. ii, 4, 5

[3] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1597–1604, 2009. ii, 4, 5

[4] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *CVPR 2011*, pp. 409–416, 2011. ii, 4

[5] J. Feng, Y. Wei, L. Tao, C. Zhang, and J. Sun, "Salient object detection by composition," in *2011 International Conference on Computer Vision*, pp. 1028–1035, 2011. ii, 2, 4

[6] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2189–2202, 2012. ii, 3, 4

[7] E. Rahtu, J. Kannala, and M. Blaschko, "Learning a category independent object detection cascade," in *2011 International Conference on Computer Vision*, pp. 1052–1059, 2011. ii, 4

[8] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, pp. 167–181, Sep 2004. 2

[9] J. Hays and A. A. Efros, "Scene completion using millions of photographs," *ACM Transactions on Graphics (SIGGRAPH 2007)*, vol. 26, no. 3, 2007. 2

[10] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, pp. 145–175, May 2001. 2

[11] D. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, pp. 1150–1157 vol.2, 1999. 2

[12] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: A database and web-based tool for image annotation," *International Journal of Computer Vision*, vol. 77, pp. 157–173, May 2008. 3

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, Jun 2010. 3