SALIENT OBJECT DETECTION USING BIPARTITE DICTIONARY

Yuna Seo, Donghoon Lee and Chang D. Yoo

Korea Advanced Institute of Science and Technology
Department of Electrical Engineering

ABSTRACT

This paper considers a bipartite dictionary based salient object detection algorithm that assigns one of two labels (object/background) to each superpixel of an image. The algorithm will iteratively find for each of the labels two dictionaries referred to as the bipartite dictionary, and the dictionaries will in turn update the labels of the superpixels based on the assumption that features of a particular label is better represented by the dictionary of its own label than by the dictionary of the other label. This iteration stops when convergence is reached, in other words, when there is no update. An objective function is formulated such that the bipartite dictionary and superpixel labels maximize inter-class reconstruction error while simultaneously minimize intra-class reconstruction error. The proposed algorithm is evaluated on the MSRA-1000 dataset. Experimental results show that the proposed algorithm performs better than state-of-the-art algorithms for the dataset when the initial conditions are set appropriately. We have also found that the proposed algorithm tends to highlight salient objects more uniformly than other algorithms.

Index Terms— saliency, salient object, iterative, sparse representation, dictionary

1. INTRODUCTION

Human brain detects informative parts in image before recognizing the image [1]. This process allows the brain to swiftly analyze a large image by highlighting important parts of the image from background. To mimic this process, many researchers in computer vision community have been actively seeking a solution to perform salient object detection which detects important parts of an image as shown in Figure 1. Salient object detection is a key preprocessing step in many of computer vision applications such as image retargeting [2, 3], content-based image retrieval [4], image segmentation [5], and object recognition [6]. This paper considers a bipartite dictionary based salient object detection algorithm that assigns one of two labels(object/background) to each superpixel of an image. The algorithm derives a dictionary for the salient object and a dictionary for the background, and henceforth, the two dictionaries will be referred to as the bipartite dictionary.

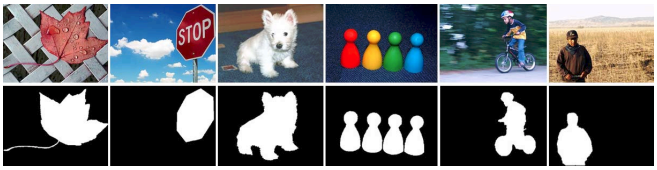


Fig. 1. The examples of salient object. Top : input image. Bottom : salient object (ground truth).

Salient object detection algorithms can be categorized into top-down and bottom-up algorithms. Top-down algorithms [7, 8] use high-level prior information about the object, and therefore, require labeled data. On the contrary, bottom-up algorithms [2, 9, 10, 11, 12, 13, 14, 15, 16, 17] are data-driven and rely on various assumptions about the properties of salient object and background.

Recently, several saliency detection algorithms based on sparse representation [8, 17] have shown promising results. A dictionary to represent background is used to determine object from background based on reconstruction error or sparsity. In [8], the saliency value of each image patch is computed depending on the sparsity of it over the entire image based on background dictionary learned from natural training images. In [17], the saliency value is defined by integrating dense and sparse reconstruction error based on a predefined background dictionary which is composed of features of the superpixels at the boundaries of a given image, and the degree of saliency or probability that a superpixel is an object is proportional to this value. In contrast to [8] in which a general dictionary is learned from many natural training images, [17] constructs a background dictionary for each individual image, and therefore, is more image-specific. However, [17] has two limitations: 1) once an initial background dictionary is heuristically determined (i.e. the boundaries of an image), it is fixed until saliency detection procedure is completely done; 2) only background dictionary is considered in this algorithm.

In this paper, a bottom-up salient object detection algorithm based on bipartite dictionary is considered. We formulate salient object detection problem as assigning each superpixel into object or background. Each superpixel assignment is initialized, and iteratively updated based on an objective function which is designed to maximize inter-class reconstruction error while minimize intra-class reconstruction

error.

The rest of the paper is organized as follows. The proposed algorithm is described in Section 2. Experimental results are shown in Section 3. Finally, conclusions are presented in Section 4.

2. THE PROPOSED METHOD

In this section, the details of the proposed algorithm are provided. First, features used in the algorithm are described. Second, a short review on sparse representation is given. Third, an extensive description on the proposed algorithm is given.

2.1. Feature

Given an image, $\mathbf{I} \in \mathbb{R}^{h \times w}$, we generate superpixels, $\mathcal{S} = \{\mathbf{s}_i\}_{i=1}^{N}$, using the Simple Linear Iterative Clustering (SLIC) algorithm [18], where N is the number of generated superpixels. Given a superpixel, $\mathbf{s}$, we extract a feature vector, $\mathbf{x} = (R, G, B, L, a, b, x, y) \in \mathbb{R}^8$, in the same manner as [17]. Here, R, G, B, L, a, b and (x, y) are respectively the mean values of pixels in $\mathbf{s}$ in RGB color space and CIELAB color space, and the mean coordinates of $\mathbf{s}$.

2.2. Sparse Representation

Given an input $\mathbf{x} \in \mathbb{R}^m$ and a dictionary $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_N] \in \mathbb{R}^{m \times N}$, sparse representation is to represent $\mathbf{x}$ as a linear combination of sparse coefficient $\boldsymbol{\alpha} \in \mathbb{R}^N$ and corresponding dictionary. Generally, the coefficient $\boldsymbol{\alpha}$ can be obtained by solving the following l_1-regularized minimization problem,

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1 , \tag{1}$$

where λ is the regularization parameter which controls the sparsity of $\boldsymbol{\alpha}$. The first term of Equation (1) represents the reconstruction error, and the second term is a regularization to impose sparsity constraint to $\boldsymbol{\alpha}$.

In this paper, Least Angle Regression (LARS) algorithm [19] implemented by [20] [1] is used to obtain $\boldsymbol{\alpha}$.

2.3. Main Algorithm

From image $\mathbf{I}$ with superpixels $\mathcal{S} = \{\mathbf{s}_i\}_{i=1}^{N}$, feature vectors $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{N}$ are extracted from each superpixel.

Let $\mathcal{O}$ and $\mathcal{B}$ be the sets of indices of superpixels labeled as object and background, respectively. From $\mathcal{O}$, the object dictionary $\mathbf{D}^O \in \mathbb{R}^{8 \times N_O}$ is constructed as the feature vectors extracted from all superpixels labeled as object. That is, $\mathbf{D}^O = [\mathbf{x}_{o_1}, \mathbf{x}_{o_2}, \ldots, \mathbf{x}_{o_{N_O}}]$ where $\mathcal{O} = \{o_1, o_2, \ldots, o_{N_O}\}$. The background dictionary $\mathbf{D}^B \in \mathbb{R}^{8 \times N_B}$ is constructed in the same manner, that is, $\mathbf{D}^B = [\mathbf{x}_{b_1}, \mathbf{x}_{b_2}, \ldots, \mathbf{x}_{b_{N_B}}]$ where $\mathcal{B} = \{b_1, b_2, \ldots, b_{N_B}\}$. Here, N_O and N_B are respectively the numbers of superpixels in object and background, and $N = N_O + N_B$.

Then, we can define reconstruction errors for i-th superpixel, $i \in \mathcal{O}$, as follows:

$$\varepsilon_i^O = \left\|\mathbf{x}_i - \mathbf{D}^{O\setminus i}\boldsymbol{\alpha}_i^O\right\|_2^2 \tag{2}$$

$$\varepsilon_i^B = \left\|\mathbf{x}_i - \mathbf{D}^B\boldsymbol{\alpha}_i^B\right\|_2^2 . \tag{3}$$

Note that since $i \in \mathcal{O}$, $\mathbf{x}_i$ is excluded from $\mathcal{O}$. Similarly, we can define reconstruction errors when $i \in \mathcal{B}$:

$$\varepsilon_i^O = \left\|\mathbf{x}_i - \mathbf{D}^O\boldsymbol{\alpha}_i^O\right\|_2^2 \tag{4}$$

$$\varepsilon_i^B = \left\|\mathbf{x}_i - \mathbf{D}^{B\setminus i}\boldsymbol{\alpha}_i^B\right\|_2^2 . \tag{5}$$

Here, reconstruction weights, $\boldsymbol{\alpha}_i^O$ and $\boldsymbol{\alpha}_i^B$, can be obtained by solving the minimization problems in Equation (1).

Equation (2) and (5) are intra-class reconstruction errors which should be minimized, while Equation (3) and Equation (4) are inter-class reconstruction errors which should be maximized. From this, we can design an objective function by combining Equation (2-5),

$$\begin{aligned}\Psi(\mathcal{O}, \mathcal{B}) = {} & \beta_B \sum_{i \in \mathcal{O}} \varepsilon_i^B + \beta_O \sum_{i \in \mathcal{B}} \varepsilon_i^O \\ & - \beta_O \sum_{i \in \mathcal{O}} \varepsilon_i^O - \beta_B \sum_{i \in \mathcal{B}} \varepsilon_i^B ,\end{aligned} \tag{6}$$

where, $\beta_B = 1/(\frac{N_O}{N} + \gamma)$ and $\beta_O = 1/(\frac{N_B}{N} + \gamma)$ shrink the reconstruction errors to resolve the unbalancing problem due to the different number of bipartite dictionary.

$\mathcal{O}$ and $\mathcal{B}$ that maximize Equation (6) are maximizing intra-class reconstruction errors and are simultaneously minimizing inter-class reconstruction errors. To find $\mathcal{O}$ and $\mathcal{B}$ that maximize Equation (6) is NP-hard problem. Alternatively, we propose an iterative algorithm that finds local maximum. The pseudo code of the proposed algorithm is described in Algorithm 1.

In the proposed algorithm, we focus on the variation of objective score when the assignment is changed. When j-th superpixel assignment is changed from $\mathcal{O}$ to $\mathcal{B}$, the objective score for sets $\mathcal{O}' = \mathcal{O}\setminus j$ and $\mathcal{B}' = \mathcal{B} \cup j$ is changed as:

$$\begin{aligned}\Psi(\mathcal{O}', \mathcal{B}') = {} & \beta_{B'} \sum_{i \in \mathcal{O}'} \varepsilon_i^{B'} + \beta_{O'} \left(\sum_{i \in \mathcal{B}} \varepsilon_i^{O'} + \varepsilon_j^O \right) \\ & - \beta_{O'} \sum_{i \in \mathcal{O}'} \varepsilon_i^{O'} - \beta_{B'} \left(\sum_{i \in \mathcal{B}} \varepsilon_i^{B'} + \varepsilon_j^B \right) .\end{aligned} \tag{7}$$

If we discard small variations of β_B, β_O and $\varepsilon_{i, i \neq j}$ due to the assignment change, then we can write $\beta_{B'} = \beta_B$, $\beta_{O'} = \beta_O$, $\varepsilon_i^{O'} = \varepsilon_i^O$ and $\varepsilon_i^{B'} = \varepsilon_i^B$ for $i \neq j$. Then the variation of objective score is

$$\Psi(\mathcal{O}', \mathcal{B}') - \Psi(\mathcal{O}, \mathcal{B}) \simeq 2(\beta_O \varepsilon_j^O - \beta_B \varepsilon_j^B). \tag{8}$$

[1] http://spams-devel.gforge.inria.fr/downloads.html

Algorithm 1 Bipartite Dictionary based Salient Object Detection (BD)

Input: Image $\mathbf{I} \in \mathbb{R}^{h \times w}$
Compute superpixels $\mathcal{S} = \{\mathbf{s}_i\}_{i=1}^{N}$ and corresponding features $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{N}$
Initialize object set $\mathcal{O}$ and background set $\mathcal{B}$
repeat
  Construct object dictionary $\mathbf{D}^O$ and background dictionary $\mathbf{D}^B$
  for $i = 1$ *to* N **do**
    if $i \in \mathcal{O}$ **then**
      $\hat{\boldsymbol{\alpha}}_i^O = \text{argmin}_{\boldsymbol{\alpha}_i^O} \left\|\mathbf{x}_i - \mathbf{D}^{O\setminus i}\boldsymbol{\alpha}_i^O\right\|_2^2 + \lambda_O \left\|\boldsymbol{\alpha}_i^O\right\|_1$
      $\hat{\boldsymbol{\alpha}}_i^B = \text{argmin}_{\boldsymbol{\alpha}_i^B} \left\|\mathbf{x}_i - \mathbf{D}^B\boldsymbol{\alpha}_i^B\right\|_2^2 + \lambda_B \left\|\boldsymbol{\alpha}_i^B\right\|_1$
      $\varepsilon_i^O = \left\|\mathbf{x}_i - \mathbf{D}^{O\setminus i}\boldsymbol{\alpha}_i^O\right\|_2^2,\ \varepsilon_i^B = \left\|\mathbf{x}_i - \mathbf{D}^B\boldsymbol{\alpha}_i^B\right\|_2^2$
      if $\beta_O\varepsilon_i^O > \beta_B\varepsilon_i^B$ **then**
        $\mathcal{O} \leftarrow \mathcal{O}\setminus i,\ \mathcal{B} \leftarrow \mathcal{B} \cup \{i\}$
      end if
    else
      $\hat{\boldsymbol{\alpha}}_i^O = \text{argmin}_{\boldsymbol{\alpha}_i^O} \left\|\mathbf{x}_i - \mathbf{D}^O\boldsymbol{\alpha}_i^O\right\|_2^2 + \lambda_O \left\|\boldsymbol{\alpha}_i^O\right\|_1$
      $\hat{\boldsymbol{\alpha}}_i^B = \text{argmin}_{\boldsymbol{\alpha}_i^B} \left\|\mathbf{x}_i - \mathbf{D}^{B\setminus i}\boldsymbol{\alpha}_i^B\right\|_2^2 + \lambda_B \left\|\boldsymbol{\alpha}_i^B\right\|_1$
      $\varepsilon_i^O = \left\|\mathbf{x}_i - \mathbf{D}^O\boldsymbol{\alpha}_i^O\right\|_2^2,\ \varepsilon_i^B = \left\|\mathbf{x}_i - \mathbf{D}^{B\setminus i}\boldsymbol{\alpha}_i^B\right\|_2^2$
      if $\beta_O\varepsilon_i^O < \beta_B\varepsilon_i^B$ **then**
        $\mathcal{O} \leftarrow \mathcal{O} \cup \{i\},\ \mathcal{B} \leftarrow \mathcal{B}\setminus i$
      end if
    end if
  end for
until $\mathcal{O}, \mathcal{B}$ don't change or reach max iteration number n
Output: Final object set $\mathcal{O}$ and background set $\mathcal{B}$

The variation of the objective score due to the assignment change can be approximately computed by directly comparing $\beta_B\varepsilon_i^B$ to $\beta_O\varepsilon_i^O$, and we can derive Algorithm 1 based on this approximation.

In the following experiments, we generate superpixels at eight different scales and then average all results to obtain a more reliable result. By assuming that the objects are more likely to be at the center of image (center prior), the superpixels located at the center can be initially labeled as object. We found importance of initialization, so we use the result of [17] as initialization in our experiments to obtain better results.

3. EXPERIMENT

We compare performance of the proposed algorithm with 12 state-of-the-art algorithms including IT [9], GB [10], SR [11], LC [12], AC [13], FT [14], CA [2], HC [15], RC [15], CB [16], DSR [17], HS [21].

3.1. Dataset

We evaluate the proposed algorithm on the MSRA-1000 dataset [14]. The MSRA-1000 dataset contains 1,000 images selected from the MSRA dataset [7] and provides ground-truth masks which are manually segmented into object and background regions.

3.2. Implementation Details

3.2.1. Experimental Setting

We generate superpixels at eight different scales varying the number of superpixels N (N = 50, 100, 150, 200, 250, 300, 350, 400) [17]. The regularization parameters λ_O and λ_B are empirically set to 0.01. The parameter γ used to compute shrinkage parameter β_O and β_B is empirically set to 0.1. The max iteration number n is set to 30, but most images converge much earlier.

3.2.2. Evaluation Metrics

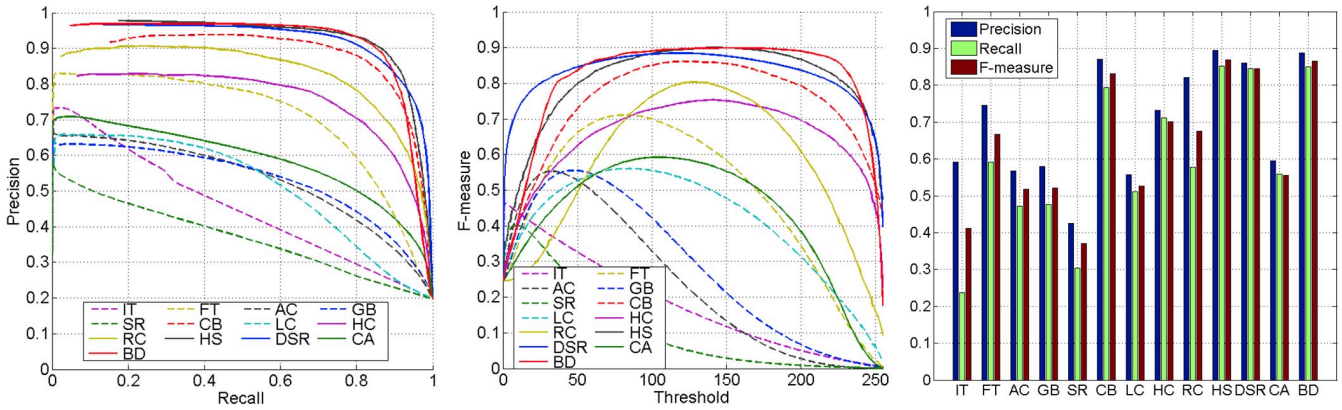
We evaluate all algorithms with respect to precision, recall and F-measure. The precision value is a ratio of the number of correctly detected pixels to the number of all selected pixels by algorithm, while the recall value is a ratio of the number of correctly detected pixels to the number of all ground-truth salient pixels. The final saliency map is binarized with a threshold in the range of 0 to 255, and then precision and recall values are computed by comparing binarized map with ground-truth mask. The precision-recall curve is obtained by averaging the precision and recall values over all images at each threshold. The F-measure represents overall performance and is computed by following equation,

$$F_\alpha = \frac{(1+\alpha^2)\text{Precision} \times \text{Recall}}{\alpha^2\text{Precision} + \text{Recall}}, \tag{9}$$

where α^2 is set to 0.3 to emphasize precision [14]. We also compare the mean precision, recall and F-measure with an adaptive threshold which is defined as twice the mean of saliency map of the image [14].

3.3. Experimental Results

The quantitative results of the proposed algorithm (BD) and state-of-the-art algorithms are illustrated in Figure 2. The precision-recall curves are shown in Figure 2 (a). The proposed algorithm performs better than most state-of-the-art algorithms and is also comparable to even very recent state-of-the-art algorithms such as HS [21]. The F-measure values at each threshold are plotted in Figure 2 (b). The proposed algorithm shows highest F-measure value in a wide range, which means that it is less sensitive to picking a certain threshold. Figure 2 (c) shows the mean precision, recall and F-measure of all algorithms with an adaptive threshold. The proposed algorithm achieves the almost highest precision and the best F-measure values.



(a) Precision-recall curve (b) F-measure (c) Adaptive thresholding

Fig. 2. The performance of the proposed algorithm (BD) compared with 12 state-of-the-art algorithms on MSRA-1000 dataset. (a) Precision-recall curve, (b) F-measure, (c) Precision, recall and F-measure value with an adaptive threshold.

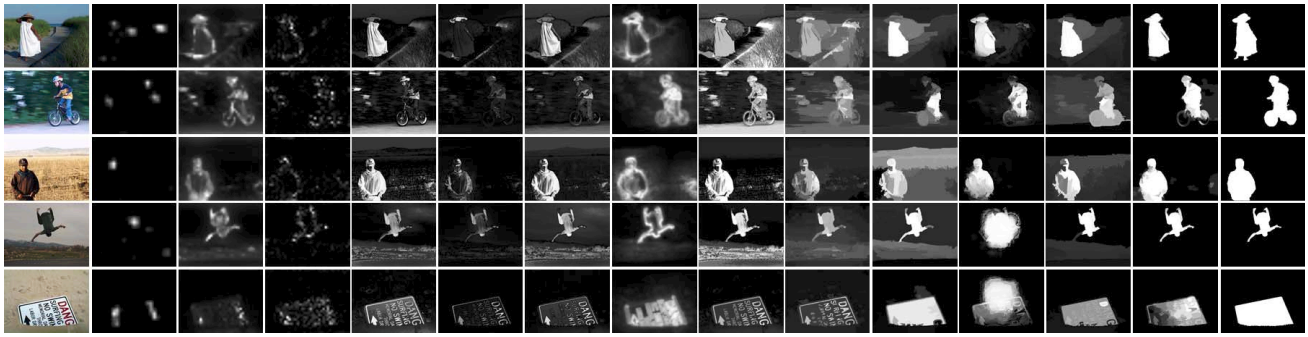


Fig. 3. Visual comparison of the proposed algorithm (BD) and 12 state-of-the-art algorithms on MSRA-1000 dataset. GT: ground truth.

Note that the goal of the proposed algorithm is to assign one of two labels to each superpixel, therefore, a result image at one scale is a binary image. After obtaining the final result (usually referred to as a saliency map) by averaging all results, each pixel in the final saliency map can have only a few discrete values. Therefore, when we plot Figure 2 (b), the curve looks like stairs. Thus, to obtain smooth curve, we temporarily compute saliency score of each superpixel using $\beta_B \varepsilon_i^B - \beta_O \varepsilon_i^O$ debasing the overall performance. $\beta_B \varepsilon_i^B - \beta_O \varepsilon_i^O$ is the approximated objective score variation due to assignment change.

Some qualitative results are also shown in Figure 3. Note that the proposed algorithm consistently generates accurate saliency map close to ground-truth and tends to highlight salient objects more uniformly than other algorithms.

4. CONCLUSION

In this paper, we have proposed the bipartite dictionary based salient object detection algorithm that assigns one of two labels (object/background) to each superpixel of an image. The algorithm iteratively find bipartite dictionary, and the dictionaries will in turn update the labels of the superpixels based on the assumption that features of a particular label is better represented by the dictionary of its own label than by the dictionary of the other label. The objective function which is designed to maximize inter-class reconstruction error and simultaneously minimize intra-class reconstruction error has been presented. Experimental results have shown that the proposed algorithm performs better than state-of-the-art algorithms when the initial conditions are set appropriately.

5. ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No.NRF-2010-0028680) and by the Industrial Strategic Technology Development Program (10044009) funded by the Ministry of Knowledge Economy (MKE, Korea)

6. REFERENCES

[1] J. K Tsotsos, S. M Culhane, W. Y. Kei Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artificial intelligence*, vol. 78, no. 1, pp. 507–545, 1995.

[2] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE PAMI*, vol. 34, no. 10, pp. 1915–1926, 2012.

[3] R. Achanta and S Susstrunk, "Saliency detection for content-aware image resizing," in *ICIP*, 2009, pp. 1005–1008.

[4] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, "Sketch2photo: internet image montage," in *ACM Transactions on Graphics*, 2009, p. 124.

[5] W. Guo, C. Xu, S. Ma, and M. Xu, "Visual attention based small object segmentation in natual images," in *ICIP*, 2010, pp. 1565–1568.

[6] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?," in *CVPR*, 2004, pp. II–37.

[7] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE PAMI*, vol. 33, no. 2, pp. 353–367, 2011.

[8] A. Borji and L. Itti, "Exploiting local and global patch rarities for saliency detection," in *CVPR*, 2012, pp. 478–485.

[9] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE PAMI*, vol. 20, no. 11, pp. 1254–1259, 1998.

[10] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in neural information processing systems*, 2006, pp. 545–552.

[11] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *CVPR.*, 2007, pp. 1–8.

[12] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *ACM international conference on Multimedia*, 2006, pp. 815–824.

[13] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk, "Salient region detection and segmentation," in *Computer Vision Systems*, pp. 66–75. 2008.

[14] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *CVPR*, 2009, pp. 1597–1604.

[15] 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.

[16] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li, "Automatic salient object segmentation based on context and shape prior," in *BMVC*, 2011, p. 7.

[17] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," 2013.

[18] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels," *École Polytechnique Fédéral de Lausssanne (EPFL), Tech. Rep*, vol. 149300, 2010.

[19] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, et al., "Least angle regression," *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[20] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *JMLR*, vol. 11, pp. 19–60, 2010.

[21] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *CVPR*, 2013.