# A Survey on Approaches for Saliency Detection with Visual Attention

Qi Zhang1,[a]

[1]School of Computer Science, Chang'an University, 710064 Xi'an, China

**Abstract.** Most existing approaches for detecting salient areas in natural scenes are based on the saliency contrast within the local context of image. Nowadays, a few approaches not only consider the difference between the foreground objects and the surrounding background areas, but also consider the saliency objects as the candidates for the center of attention from the human's perspective. This article provides a survey of saliency detection with visual attention, which exploit visual cues of foreground salient areas, visual attention based on saliency map, and deep learning based saliency detection. The published works are explained and descripted in detail, and some related key benchmark datasets are briefly presented. In this article, all documents are published from 2013 to 2018, giving an overview of the progress of the field of saliency detection.

## 1 INTRODUCTION

Saliency detection is the burning topic in computer vision that has been applied to many tasks, such as image classification and retrieval, semantic segmentation, object recognition [1]. Image and video saliency detection is a sophisticated but valuable issue in many fields for real-world applications due to the difficulty of manually monitor large amounts of data from vision sensors, intelligent video surveillance is greatly worthy of study. Thus, researchers are increasingly transferring the attention to the perception of human action recognition, and trying to apply it to human-computer interaction.

Significant progress has been made, but the detection of conspicuous objects remains a difficult issue. Most conventional detecting methods for conspicuous objects are based on training to detect a specific target category by the difference of saliency in the local context of image [2]. Human vision is a clustered visual scene that allows us to focus on common and prominent objects quickly, and saliency detection has attracted the attention of researchers and gotten much significant progress. However, it remains challenging problems, such as distinct patterns, spatial compactness, and objectness measure [3].

Existing algorithms on saliency detection in general can be divided into three core components: visual cues of foreground salient areas, visual attention based on saliency map, and deep learning based saliency detection. The visual cues of foreground salient areas are mainly extracted with some features (such as color, distinct pattern, illumination intensity, view orientation), and some prior knowledge (such contrast, spatial compactness, smooth appearance, uniqueness, boundary). It is known that these visual cues draw our attention [4]. After viewing the image in a while, the human can observe multiple areas of interest and various typical salient point through the image content. Deep learning based saliency detection methods, e.g. CNNs, capture typical high-level features to detect salient objects that are prominent in a particular size and category. In sum, saliency detection performance based on deep learning is generally much better than other traditional bottom-up models [1].

From the above-mentioned works, the first two components of background areas are much similar to the image limitations between the local regions and the global regions usually. Specifically, the foreground regions are similar with visual cures. For example, Cheng et al. [5] propose a projecting saliency detection based on the spatial coherence of the surrounding areas and the contrast of the global regions at same time. Wei et al. [6] address the saliency detection with two features, including the restrictions and the connections of background areas. Wang et al. [7] measure the saliency by propagating a label that yields the initial labeling of non-labeled elements according to the affinity of the pairs. A low-level saliency index is useful for the simple images and videos, but not necessarily robust in difficult scenarios. For the saliency prediction, more high-level image information and the context should be taken into account [8]. In recent years, CNNs have also achieved the top and advanced performance for saliency detection issues. Li et al. [9] extract the multi-scale features first and forms a fully connected regression network to infer the saliency score for the patch of each segment. Wang et al. [10] train the DNN-L and a DNN-G network using the local patch functions and the global candidate features to measure salience. Wang et al. [11] introduce prior knowledge to the complete and recurrent fully convolutional network for precise saliency reasoning.

In this paper, we describe the advanced methods about saliency detection in recent years. All the publications in question are reported from 2013 to 2018. Most of them are from CVPR, ECCV, PAMI, TIP, etc., and some recent documents are just filed in arXiv as well. Specifically, the focus points of this paper are the methods based on deep learning and visual attention. The details of the related methods are described in following sections, and the rest of the paper is organized as follows. Section 2 shows some datasets in the experiments. Section 3 examines the deep learning models in saliency detection. Section 4 describes the visual attention based saliency detection. Section 5 presents the conclusions and some future works.

## 2 BENCHMARK DATASETS AND EVALUATION CRITERIONS

### 2.1. Benchmark datasets

There are nine common benchmark datasets for the saliency detection problems, including MSRA [12], THUS [5], DUT-OMRON [3], MIT300 [13], ECSSD [14], PASCAL-S [15], SOD [16], HKU-IS [17], and SED1 [18]. The MSRA dataset [12] contains 5,000 images with prominent bounding boxes of terrestrial ground truth salient regions. The subset of MSRA dataset is titiled ASD dataset, which consists of 1,000 images in the segmentation mask of ground truth pixels, i.e., namely the salient objects in each image with exact human label. The THUS dataset [5] contains 10,000 image, each of which is annotated in the area of ground truths for remarkable salience objects. The DUT-OMRON dataset [3] includes 5,168 images labeled with multiple objects at different locations and scales in the cluttered backgrounds. The MIT300 dataset [13] is composed of 300 images with eye tracking data from 39 observers for ocular fixation predictions. The ECSSD dataset [14] contains 1,000 structurally complex images, while the PASCAL-S dataset [15] consists of 850 natural images as one of the most difficult saliency datasets. The SOD dataset [16] is another challenging one that contains 300 images from the Berkeley segmentation dataset. In this dataset, there is an image that makes up multiple instances of objects of different size. The HKU-IS dataset [17] contains 4,447 images which are divided into training and testing examples. The SED1 dataset [18] consists of 100 images with different-sizes and different-locations objects. Some examples of input images and salient object ground truths from above-mentioned datasets are illustrated in Figure 1, and the overview description of datasets are shown in Table 1.

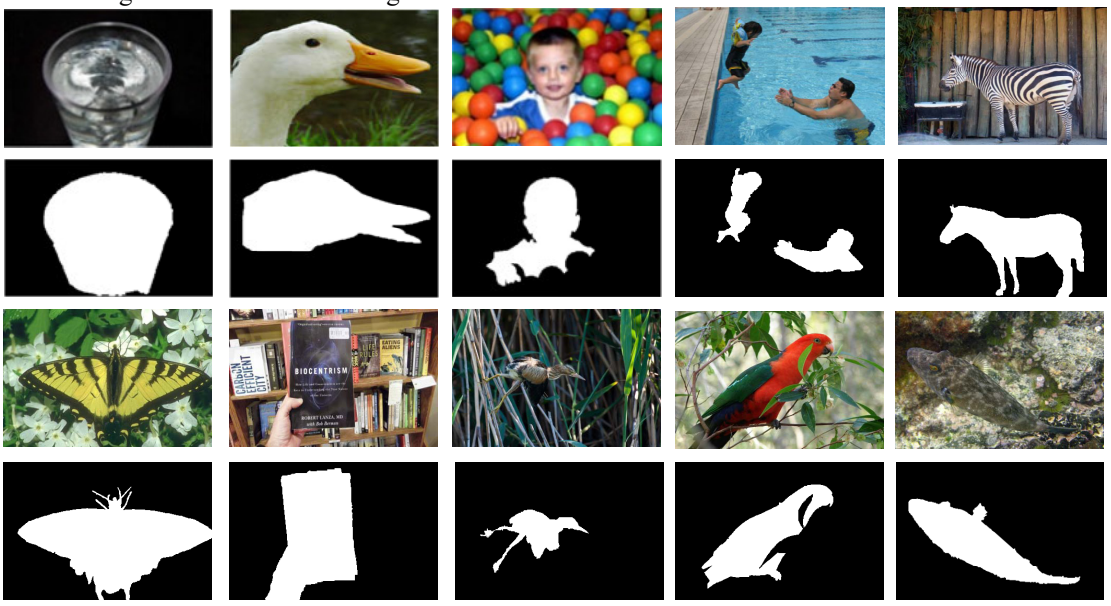


**Figure 1.** Examples of input images and salient object ground truths from datasets.

**Table 1.** The description of nine datasets for saliency detection.

| dataset | # of images | issues | format |
|---|---|---|---|
| MSRA | 5,000 | N/A | JPEG |
| THUS | 10,000 | region ground truth | JPEG |
| DUT-OMRON | 5,168 | multiple objects at different locations and scales in the cluttered background | JPEG |
| MIT300 | 300 | N/A | JPEG/BMP |
| ECSSD | 1,000 | structurally complex image | JPEG |
| PASCAL-S | 850 | natural image | JPEG |
| SOD | 300 | Multiple objects with different sizes | JPEG |
| HKU-S | 4,447 | N/A | JPEG/BMP |
| SED | 100 | Largely different sizes and locations | JPEG |

### 2.2. Evaluation criterions

Six general quantitative criterions in the evaluation of saliency detection are detailed as follows, such precision and recall (PR) curve, F-measure, and area under curve (AUC), mean absolute error (MAE), normalized scan-path saliency (NSS), and similarity (S). In the PR curve, the precision rate at y-coordinate is defined as a pronounced pixel accuracy that has been correctly retrieved for all pixels in the extracted region, and the recall rate at x-coordinate indicates the ratio of accurately retrieved salient pixels to all pixels in ground truth of saliency pixels. Most of the works segment the saliency map with the threshold ranging from 0 to 1 at 0.05 interval, and compute the precision and recall at each value of the threshold to plot the PR curve. The F-measure is the overall performance measurement, which is calculated by the precision and recall with weighted harmonics [1]. According to the PR curve and F-measure, the AUC score is computed by the area under the plotted PR curve with the positive rates, including the true positive rates and the false positive rates. The MAE represents the mean difference between the saliency map and the ground truth [3]. The NSS metric is computed by the means of an average standardized salience value across all fixation positions of the image [19]. The S score metric is defined to measure the probability distributions of the fixed saliency maps from the intersections of the histogram [4]. In Table 2, we summarize these evaluation criterions.

**Table 2.** Summary of different metrics in saliency evaluation. ↑ represents that the metric value is larger for the better saliency results, while ↓ denotes that the smaller metric value is for the better saliency results.

| metric | PR curve | F-measure | AUC | MAE | NSS | S |
|---|---|---|---|---|---|---|
| value | N/A | ↑ | ↑ | ↓ | ↑ | ↑ |

## 3 DEEP LEARNING BASED SALIENCY DETECTION

Recently, in the task of computer vision, deep convolutional neural networks (CNN) has reached the forefront of performance and has become a powerful method for extracting high-level semantic features in various scales instead of building the conventional hand-craft features. The standard feed-forward CNN and the recently proposed networks are used by grouping of convolutional, pooling, and fully connected layers. In this type of methods, a fixed spatial size image is taken as input, then the convolutional and pooling layers are designed to control the capacity of the model to increase the size of the receptive field, which gives an expression of a rough and high semantic representations. Most of these frameworks are trained from end to end automatically with a stochastic back propagation algorithm.

Many saliency detection methods based on deep learning have sprung out and obtained the performance near the human-level on benchmark datasets, which will be presented in the following paragraphs. Li et al. [19] make the first effort to train a deep method for saliency detection without any human annotation, in which the important points are led by the supervised fusion. In other words, it produces the weak but fast fusion process of the unsupervised models for saliency detection with useful supervisory signals. At this point, two streams of image fusion are combined with the intra-image fusion and the inter-image fusion, in order to produce the learning curriculum and pseudo ground-truth for overseeing and training the deep models for salient object detection. Wang et al. [20] train a novel ranking saliency detection method, which utilizes the region-based convolutional neural network (R-CNN) features and deep learning based object proposal. To solve the ranking saliency problem, the constraint is convinced that most of positive samples have the higher scores than the negative ones. Due to the large size of the deep model parameters and the low amount of the training data, the classification in the primitive space and ranking process are not optimal. The novel ranking and kernelized model in the subspace is proposed by jointly learning, which contains a subspace projection and a ranking SVM classifier. The projections are intended to measure a pair of distances in the low-dimensional space, while the ranking score for each proposition is assigned by the learned ranking sequences for an image. Finally, the saliency map is computed by the fusion of weighted good ranked candidates. Kim et al. [21] propose that CNN can be used as a classifier for multiple categories for the objection proposals in the image regions and predicting the binary maps accordingly. In this remarkable method, two CNNs are combined to collect the global and local saliency information together, which predicts the shape regions in the CNN-driven form considering not only the intermediate information but also the small-region information. Furthermore, after learning the shape features of a foreground object with CNN, the proposed model from Kim et al. estimates the complete saliency map of the target area and then refines the rough saliency map using information from a certain low level to a specific middle level. In this way, the typical architecture of low resolution feature maps remain unchanged at whatever level of pixel variability, which helps to retrieve the specific information at the object level. By the experimental comparison and validation, this method is useful for dense segmentation of foreground objects in the task of saliency detection.

Recurrent neural networks (RNN) and fully connect networks (FCN) are also applied in the end-to-end densely segmenting saliency detection [1], [11], [22]. Wang et al. use the prior knowledge of saliency and incorporate it into the networks to automatically refine the current saliency maps with the context information. Liu et al. [22] propose the preceding saliency maps using large amount of low-level characteristics, which makes it difficult to learn the best multi-scale information across the entire network, and further introduce a hierarchical RNN to refine the saliency maps by integrating local information of image contexts. Zhang et al. [1] incorporate the extracted features from

FCN into the transition probability and the absorption of the Markov chain studied in semantic segmentation for saliency detection. A little connected graph that captures all context information for boundary nodes is built in the proposed method and treated as the transient nodes and absorbing nodes of the absorbent Markov chain. These nodes are used to represent the significant values for calculating the weights of the spatial paths and coordinates. The different hierarchies of deep features extracted by RNN are encoded in the learned transition probability matrix. In order to improve the performance of deep saliency detection methods combing RNN and FCN, some techniques are considered and investigated to improve the saliency maps, such as the ordering of pairs, angular coupling, and so on. The advantage of these methods lies on its high accuracy and less computational time without requiring any preprocessing or a priori knowledge on salient capture.

## 4 VISUAL ATTENTION BASED SALIENCY DETECTION

Visual attention has a lot of progress in some computer vision tasks. Instead of constructing features from the whole regions, visual attention based methods extracts the iteratively refined saliency maps on arbitrary-size image sub-regions. These approaches, in general, can achieve better performance from the large annotated training data, which is more promising by the expression of beneficial features and hidden patterns. Specifically, Xu et al. [23] present the saliency detection method combining cognitive-based objectness and image-based saliency to lessen negligible background information. By introducing the downward attention priors, a computational selective attention model is proposed for object segmentation on the saliency map. Inspired by the visual avian pathways, Wang et al. [24] design a multi-level visual attention model for the task of saliency detection, especially the atomic nuclei linked to the visual attention mechanism. In the first hierarchy, the self-information is computed for the primary saliency maps using the optic tectum neuron responses. In the second hierarchy, the tecto-isthmal projection of the centrifugal pathway is simulated and estimated with regularized random walks ranking.

Another important works of saliency modeling show that human visual attention tends to be focused in the center of images, since the super-pixels of projecting maps are located far from the center, even near the boundary of the images as more as possible. There are two classes of knowledge governing the assignment of human's attention, including with upward and downward attention. Jian et al. [25] first propose a concept of visual attention-aware and build a human visual model system. The informative and directional patches of images are treated as visual stimuli and neuronal cues for humans to interpret and detect salient objects. These two significant patches are individually extracted and in parallel learned from the channel of both the discriminant color and the light intensity. In addition, an improved and extended wavelet-based salient region detecting method is utilized to extract the beneficial patches full of visually information. Kuen et al. [26] suggest a recurrent and regressive attention network based on the convolution and deconvolution layers, which can locally improve the saliency maps of a progressively selected regions in the small size. In this work, the center surrounding approach is extended for images to video by adding another dimension.

Besides, Wang et al. [8] propose an increase augment in the feed forward neural networks that is a new pyramidal grouping of pooling modules and a sophisticated a multi-step refinement mechanism for saliency detection. More specifically, in the works of Wang et al., the deep feed forward network is designed to generate and produce the raw forecast salience maps with a much more detailed structure loss, then the refinement net is integrated with the context information of local patches in the image for the previous maps in the scene, and finally the pyramid pooling is applied for the global context aggregation based on different regions at multiple stages. This work is compared to the experimental evaluations over the benchmark datasets for all the published methods. The interested reader may refer to some more survey publications on the research of saliency detection [27] [28].

## 5 CONCLUSIONS

With the development of saliency detection application, it becomes more and more widespread such as object detection, image segmentation, and intelligent video surveillance, but it is still a difficult task in the field of computer vision. Recently, a research trend about saliency detection is rapidly popular to figure out the issues by learning hierarchical nonlinear function and hierarchical relationships from the established deep learning based and visual attention based methods, and making a lot of progress at several points. A large majority of conventional methods construct the filters and build hand-crafted features for recognition, which relies on the segmentation up or down with large cost of time consuming and low quality of inefficient learning. To address the diversity and deficiency of hand-crafted features, the methods with deep learning architecture are driven by data, which can automatically extract functionality from the original data. Two typical kinds of deep learning based methods are devoted to extract spatial and temporal features, including convolutional neural network (CNN) and recurrent neural network (RNN). Instead of extracting features across the whole regions, the visual attention based methods extracts the iteratively refined saliency maps on arbitrary-size image sub-regions. These approaches are generally promising performance with good representations of informative features and hidden patterns obtained from the large size of annotated training instruction data. There are nine common benchmark datasets and six evaluation criterions. As the accuracy is improved, the significant in practice is boosted gradually to make the task of saliency detection more and more active.

## REFERENCES

1. Zhang, L., Ai, J., Jiang, B., Lu, H., & Li, X. (2018). Saliency Detection via Absorbing Markov Chain With Learnt Transition Probability. IEEE Transactions on Image Processing, 27(2), 987–998.
2. Guo, C., Ma, Q., & Zhang, L. (2008). Spatio-temporal Saliency detection using phase spectrum of quaternion fourier transform. In 2008 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1–8).
3. Zhang, L., Yang, C., Lu, H., Ruan, X., & Yang, M.-H. (2017). Ranking Saliency. IEEE Transactions on Pattern Analysis and Machine Intelligence.
4. Judd, T. M., Ehinger, K. A., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In 2009 IEEE 12th International Conference on Computer Vision (pp. 2106–2113).
5. Cheng, M.-M., Zhang, G.-X., Mitra, N. J., Huang, X., & Hu, S.-M. (2011). Global contrast based salient region detection. In IEEE Transactions on Pattern Analysis and Machine Intelligence (Vol. 37, pp. 409–416).
6. Wei, Y., Wen, F., Zhu, W., & Sun, J. (2012). Geodesic saliency using background priors. In ECCV'12 Proceedings of the 12th European conference on Computer Vision - Volume Part III(pp. 29–42).
7. Yang, C., Zhang, L., Lu, H., Ruan, X., & Yang, M.-H. (2013). Saliency Detection via Graph-Based Manifold Ranking. In 2013 IEEE Conference on Computer Vision and Pattern Recognition (pp. 3166–3173).
8. Wang, T., Borji, A., Zhang, L., Zhang, P., & Lu, H. (2017). A Stagewise Refinement Model for Detecting Salient Objects in Images. In 2017 IEEE International Conference on Computer Vision (ICCV) (pp. 4039–4048).
9. Li, G., & Yu, Y. (2015). Visual saliency based on multiscale deep features. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 5455–5463).
10. Wang, L., Lu, H., Ruan, X., & Yang, M.-H. (2015). Deep networks for saliency detection via local estimation and global search. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3183–3192).
11. Wang, L., Wang, L., Lu, H., Zhang, P., & Ruan, X. (2016). Saliency Detection with Recurrent Fully Convolutional Networks. In European Conference on Computer Vision (pp. 825–841).
12. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., & Shum, H.-Y. (2011). Learning to Detect a Salient Object. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(2), 353–367.
13. Judd, T., Durand, F., & Torralba, A. (2012). A Benchmark of Computational Models of Saliency to Predict Human Fixations.
14. Yan, Q., Xu, L., Shi, J., & Jia, J. (2013). Hierarchical Saliency Detection. In 2013 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1155–1162).
15. Li, Y., Hou, X., Koch, C., Rehg, J. M., & Yuille, A. L. (2014). The Secrets of Salient Object Segmentation. In CVPR '14 Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (pp. 280–287).
16. Movahedi, V., & Elder, J. H. (2010). Design and perceptual validation of performance measures for salient object segmentation. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops (pp. 49–56).
17. Li, G., & Yu, Y. (2015). Visual saliency based on multiscale deep features. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 5455–5463).
18. Alpert, S., Galun, M., Brandt, A., & Basri, R. (2012). Image Segmentation by Probabilistic Bottom-Up Aggregation and Cue Integration. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(2), 315–327.
19. Xie, L., Pan, W., Tang, C., & Hu, H. (2014). A pyramidal deep learning architecture for human action recognition. International Journal of Modelling, Identification and Control, 21(2), 139–146.
20. Wang, T., Zhang, L., Lu, H., Sun, C., & Qi, J. (2016). Kernelized Subspace Ranking for Saliency Detection. In European Conference on Computer Vision (pp. 450–466).
21. Kim, J., & Pavlovic, V. (2016). A Shape-based Approach for Salient Object Detection Using Deep Learning. In European Conference on Computer Vision (pp. 455–470).
22. Liu, N., & Han, J. (2016). DHSNet: Deep Hierarchical Saliency Network for Salient Object Detection. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 678–686).
23. Xu, Y., Li, J., Chen, J., Shen, G., & Gao, Y. (2017). A novel approach for visual Saliency detection and segmentation based on objectness and top-down attention. In 2017 2nd International Conference on Image, Vision and Computing (ICIVC) (pp. 361–365).
24. Wang, X., & Duan, H. (2017). Hierarchical visual attention model for saliency detection inspired by avian visual pathways. IEEE/CAA Journal of Automatica Sinica, 1–13.
25. Jian, M., Lam, K.-M., Dong, J., & Shen, L. (2015). Visual-Patch-Attention-Aware Saliency Detection. IEEE Transactions on Systems, Man, and Cybernetics, 45(8), 1575–1586.
26. Kuen, J., Wang, Z., & Wang, G. (2016). Recurrent Attentional Networks for Saliency Detection. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3668–3677).
27. Borji, A., Cheng, M., Jiang, H., & Li, J. (2014). Salient Object Detection: A Survey. ArXiv Preprint ArXiv:1411.5878.
28. Sebastian, E., & Daniel, N. (2017). A Survey on Various Saliency Detection Methods. International Journal of Computer Applications, 161(5), 5–8.