How far we away from a perfect visual saliency detection - DUT-OMRON: a new benchmark dataset

Xiang Ruan*
Na Tong**
Huchuan Lu**

Visual saliency detection has gained more and more attentions from academic and industrial researchers in the last 3 or 4 years. Due to great efforts being put into this field, many recent proposed algorithms have very good evaluation results on existing datasets. However, we argue that visual saliency detection is still far away from perfect because such good results are mainly due to the simplicity and bias of existing datasets. In this paper, we propose a new DUT-OMRON dataset which, to our best knowledge, is the first visual saliency detection dataset that has both the bounding box and eye fixations ground-truth in large scale. We evaluated 14 state-of-the-art methods on the proposed dataset, and the accuracy curves on proposed dataset are much lower than that on existing datasets. We believe our dataset is more challenging than existing ones and therefore leave more space for researchers to improve their algorithms.

Keywords: visual saliency, dataset, fixations, bounding box

1. Introduction

As an important step toward to understand human emotion and behavior, visual saliency detection research has gained more and more attentions from industrial and academic researchers. Number of relevant papers published in top conferences and journals of computer vision has significantly increased in the last 3 or 4 years.

Eye fixation prediction and salient object detection are two major research directions of visual saliency detection. General speaking, neural computing community is more focusing on fixation prediction, but computer vision researchers are more interested in salient object detection due to its close connection to many relevant research topics like image segmentation, object detection and so on. The difference between these two research directions is not significant, some of the algorithms behind them share similar methodologies, moreover the objective of the two topics is the same, that is to predict where humans look at an image.

Algorithms proposed in the field can be roughly divided into bottom-up and top-down categories. Bottom-up methods are built by modeling hypothesis of how a region would be salient to human eyes, such hypothesis, for example, are "high contrast", "center bias", "large area", etc. Top-down methods try to address the problem from a global viewpoint that usually lead to a supervised learning framework. To name some of many, paper [1], [2], [3], [4], [5], [6] are bottom-up methods and paper [7], [8] are top-down methods.

Unlike face detection, object recognition or other computer vision problems, evaluation of saliency detection result is not straightforward as one might have expected. It is mainly because that "visual saliency" can not be clearly defined. For an given image, different observer might look at different region of the image, such variation should be taken into account in evaluation process. To date, several benchmark datasets have been proposed by researchers, and some have already become de facto standard. If we look at recent literals, evaluation curves on popular datasets are very good. In particular, Precision and Recall (P-R) curve of our recent work [9] is even close to that of human being. The precision, recall and F-measure values are all around 0.9.

However, even for the most state-of-the-art algorithms, if we test it on natural images, such as photos of a personal album, the accuracy is always not satisfactory. We argue that although benchmark result of recent works are quit good, visual saliency technology itself is far away from perfect. We see this problem is partly due to the lack of a challenging benchmark dataset.

Researchers have already built some good benchmark datasets with large number of images and reasonable evaluation metrics. However, with great progress in this field in recent years, existing datasets are no longer challenging due to many reasons. To address the problem, in this paper, we propose a new visual saliency detection dataset: DUT-OMRON dataset†. The dataset has over $5,000$ images selected from SUN [10] database with large visual variation. To our best knowledge, the dataset is the first dataset which has both bounding box and fixation points ground-truth. We evaluated various state-of-the-art algorithms using the two kinds of ground-truth respectively. Evaluation results show that our dataset is more challenging than existing ones but remain reason-

* OMRON Corporation
** Dalian University of Technology

† http://ice.dlut.edu.cn/lu/DUT-OMRON/Homepage.htm

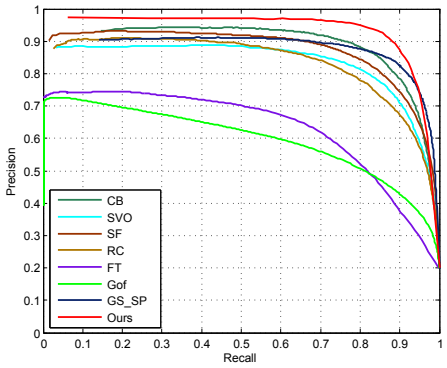

Fig. 1. Evaluation of various algorithms on MSRA-1000 dataset

able to visual saliency detection.

We discuss problems of existing dataset in Section 2 and interpret details of the proposed dataset in Section 3. Future works are concluded in Section 4.

2. Existing Datasets

2.1 Overview of Existing Datasets There is no standard benchmark dataset for visual saliency detection. Based on different methodologies, research objectives, researchers tend to present their own dataset in paper to promote their proposed algorithms. However, some datasets become more and more popular in the last 3 or 4 years because of its large number of images and well defined evaluation metrics.

Ali Borji, et al [11] presented a good summary of popular benchmark datasets. The most popular dataset for salient object detection may be MSRA [7] which includes two parts with 20,000 images and 5,000 image respectively. ASD [1] was proposed as a refined dataset using images selected from MSRA.

For fixation prediction, datasets like MIT [7], NUSEF [12], Toronto [13], Kootstra [14] are well used. We notice that there is no large scale dataset for fixation prediction, most of the datasets have less than 1,000 images.

As shown in our [9] and Borji's paper [11], current algorithms have very good evaluation results on existing datasets. Fig.1, an example from our paper [9], shows a comparison result of various algorithms on MSRA-1000 dataset. P-R curve of our algorithm is already close to human being's, leaving little space for further improvement of algorithm.

2.2 Problems of Existing Dataset However, in real applications, even the most state-of-the-art visual saliency detection technology is still not accurate. The gap between good benchmark result and less practicality for real application come from simplicity and bias of existing dataset. We summarize the problems of existing datasets as following:

- data selection problem: though popular datasets have large number of images, most of the images only have single object. Moreover, such objects are almost in the center of image with very high contrast to background. Fig.2 shows some example images of MSRA. In a real application, however, salient region will not guarantee to be in the center, and salient regions might have similar appearance or color to its surroundings.
- labeling problem: For salient object detection, ground-truth is represented by a bounding box of foreground object, on the other hand, ground-truth for fixation prediction is point set obtained by eye tracker devices. These two kinds of ground-truth are complementary to each other. bounding box is more suitable than point set for real applications or as a pre-processing of image segmentation, object detection, etc. However, manual labeling introduces some semantic bias from operators. For example, operators might pay more attention to human than other objects in images, or put more emphasis on larger objects than small objects. Such semantic bias can be eliminated by using eye tracker devices because such devices can capture operators' unconscious gazing behaviors. Unfortunately there is no existing dataset that provides both of these two ground-truth data.

Fig. 2. Some examples of MSRA dataset

3. DUT-OMRON Dataset

To address the problems mentioned above, we propose a new visual saliency detection dataset: DUT-OMRON dataset.

3.1 Data Selection We carefully selected 5,172 images from SUN dataset [10]. SUN dataset is a famous public benchmark dataset for scene recognition and object detection which has over 130,000 images. We first randomly picked up 10,000 images from SUN database and then removed images not satisfying the following criteria from the candidates:

- image is not a pure landscape image
- image has larger resolution than VGA
- image has an apparent foreground

The remained 5,172 images cover large variation of scene categories that we believe they are very similar to photos come from a personal albums. Meanwhile these images all have some regions salient to human eye, though which region is the most salient will vary to different observers. We finally normalized the images to size of $400 \times X$ or $X \times 400$, where $X \leq 400$.

We didn't set any limit on number of objects, how background should be or location of foreground objects in the image. We try to use as less criteria as possible for image selection, especially to avoid influence of hypothesis of what salient region should be. As shown in Fig. 3, our images have various contents and foreground objects.

Fig. 3. Some examples of proposed dataset

3.2 Ground-truth To our best knowledge, DUT-OMRON dataset is the first dataset that provides both bounding box and fixation point ground-truth. It should be also noticed that our dataset is made for benchmark of visual saliency detection, thus, we don't provide pixel-wise segmentation ground-truth. Actually in our opinion, saliency detection is relevant to image segmentation but not the same research topic. The goal of visual saliency detection is trying to understand where humans look at an image, but segmentation technology concerns how human eyes see different groups of visual contents. So it is obvious that different purpose of researches should use different evaluation format.

3.2.1 Bounding Box Ground-truth To better represent variation of different observers, we require 5 operators from total 25 operators to label the bounding box for each image. Unlike existing datasets, operators are asked to label multiple objects or regions that they think are salient. The number of such objects or regions is totally up to operators. We therefore obtain at least five rectangles for each image as shown in the second row of Fig.4. The bounding box ground-truth is defined by average of five operators' binary masks. Although the final bounding box is gray level (see the third row of Fig.4), for easy processing, we simply set threshold 0.5 to generate binary mask during evaluation.

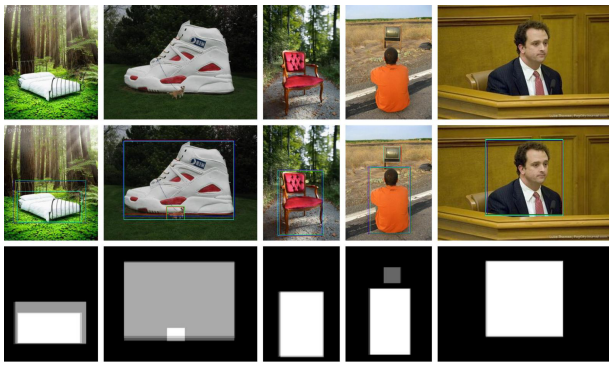

Fig. 4. Bounding box ground-truth. From top to bottom: original image, bounding boxes of five operators, average of the five binary masks

3.2.2 Fixation Ground-truth It is more complicate to generate fixation ground-truth than bounding box. We use Tobbi X1 Light Eye tracker to record operators' gazing positions. Operators sit before a monitor on which images are displayed in every two seconds without intervals. Just as labeling bounding box ground-truth, each image has data of five operators. However, due to many reasons, the raw data recorded by eye tracker has many outliers, we take the following steps to remove noise:

- delete the first fixation data to avoid the influence of center bias (people prone to look at the center of image when image suddenly shows up).
- combining with bounding box ground-truth, we remove fixations which are not in any operators' labeled bounding boxes.
- divide the fixations into three clusters by k-means since there are more than one objects in most of the images.
- only select first 90% of the fixations which has closer Euclidean distance to its cluster center.

After the removal of outliers, over 95% of all the images have more than 50 eye fixations. For the whole dataset, there are 153 fixations on average for each image. Fig.5 shows how our outliers removal scheme refines original data captured by eye tracker.

Fig. 5. Fixation ground-truth. From top to bottom: original eye-fixations, bounding box ground-truth and the eye-fixations after removing outliers

3.3 Evaluation

3.3.1 Evaluation Metrics For bounding box ground-truth, we follow conventional methods of using P-R curves and F-measure as evaluation metrics. Because P-R curve is used by many researches, it makes us easy to compare evaluation result on our dataset with that on existing datasets.

For fixation ground-truth, we first generate a saliency map by using fixations, that is to generate multiple Gaussian maps located at points of each fixation and then sum up all the maps. The final saliency map was normalized to [0..1] as shown in the second row of Fig.6. We used similar method as T. Judd, et al's [15] to utilize ROC curve as evaluation metrics. The only difference between our method and Judd's is that we set gray level 0.1 as threshold to generate binary mask (the third row of Fig.6), while Judd take the top $n\%$ of the image to obtain saliency region.

3.3.2 Evaluation Using the Dataset To show the advantages, we made evaluations on both of bounding box and fixation ground-truth of the proposed dataset. We evaluated 14 state-of-the-art algorithms ([16], [18], [19], RC of [4], [20], [3], [21], [1], [22], [17], HC of [4], [5], [6], [23]) using bounding box ground-truth, the P-R curves are shown in Fig.7. All of the curves are much

Fig. 6. Generating binary mask of fixation ground-truth. From top to bottom: fixation ground-truth, saliency map and the binary mask

lower than human's perfect curve. For a fair comparison, we show Fig.8 and Fig.9 from our paper [9], in which we evaluated the same six methods ([16], [3], [17], HC of [4], RC of [4] and ours [9]) on MSRA and the proposed dataset respectively. † It is obvious that accuracy curves in Fig.9 are much lower than that in Fig.8. Such results demonstrate that the proposed dataset is more challenging than existing ones and leaves much space for further improvement of visual saliency detection technologies.

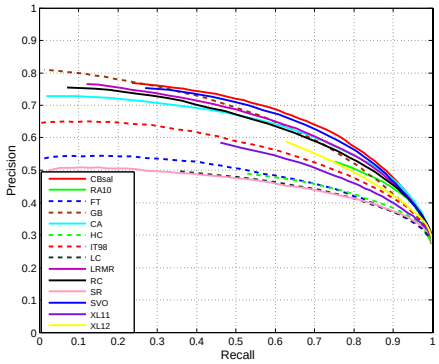

Fig. 7. Evaluations of 14 state-of-the-art methods on proposed dataset

Fig.10 is evaluation using fixation ground-truth. We utilized data from one of the 25 operators as human result to compare with three state-of-the-art fixation prediction algorithms.

The big gaps between algorithms' curves and human's in both two evaluation results give a clear message that visual saliency detection is still far away from perfect.

3.3.3 Some Analysis We also learned some interesting knowledge by analyzing the dataset.

The first one is about center bias. We combined all the 5,172 saliency maps generated by fixation ground-truth (see Section 3.3.1) and normalized it to [0..1]. Such a unified map as shown in Fig.11 shows that human fixations have strong bias to be close to the center of the

† Different from Fig.1 using MSRA-1000, Fig.8 is evaluation results on larger version of MSRA, so the accuracy is little bit worse than Fig.1

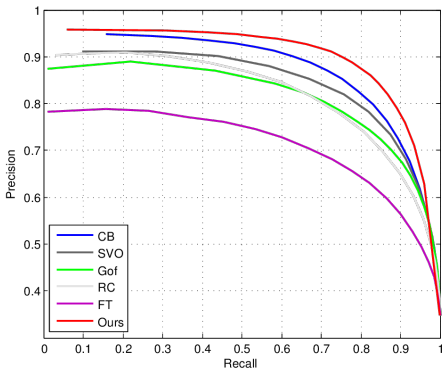

Fig. 8. Six methods including ours evaluated on MSRA dataset

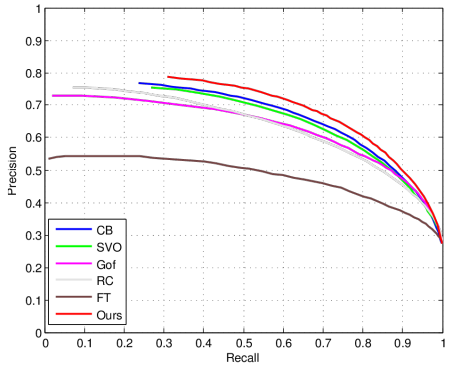

Fig. 9. Six methods including ours evaluated on proposed dataset

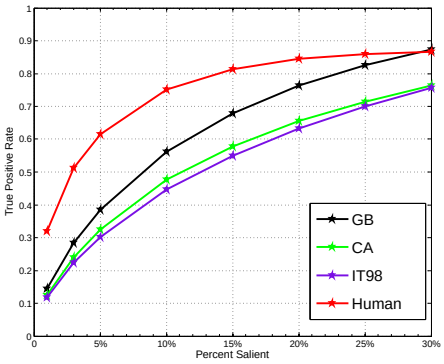

Fig. 10. Evaluations using fixation ground-truth

image. Just as mentioned above, we don't set any limit on object location during selecting images, so this result can be regarded as an evidence of so called "center bias" hypothesis.

We also checked variation of different operators. This is simply done by comparing five operators' labeling data. For each operator, we evaluated his data using ground-truth generated by other four operators. As shown in Fig.12, operators have similar ROC curves to each other. The data demonstrate the consistency of human's gazing behavior even there is variation among different individuals. It also shows that our dataset is not only challenging but also reasonable for visual saliency detection.

Fig. 11. Unified saliency map shows center bias

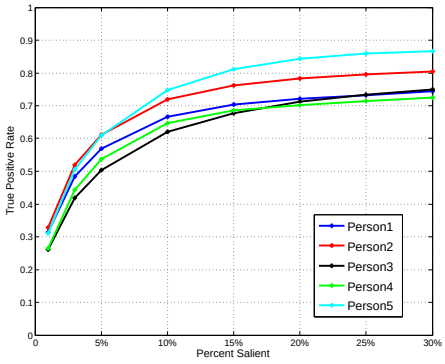

Fig. 12. Variation of different operators

4. Future Work

In this paper, we propose a new visual saliency detection benchmark dataset. Our dataset has both the bounding box and fixation ground-truth. Evaluation results indict that the proposed dataset is more challenging than existing datasets.

In the future work

- we will define an unified evaluation metrics taking both bounding box and fixation ground-truth into account.
- we will add interactive interface on dataset website to let researchers easily upload their source code or evaluation results. We will also release some automatic evaluation tools so that researchers can use it for comparing their work with other algorithms on our dataset.

.

References

(1) R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk:"Frequency tuned salient region detection", CVPR (2009)

(2) N. Bruce and J. Tsotsos:"Saliency based on information maximization", NIPS (2005)

(3) K. Y. Chang, T. L. Liu, H. T. Chen, and S. H. Lai:"Fusing generic objectness and visual saliency for salient object detection", ICCV (2011)

(4) M. M. Cheng, G. X. Zhang, N. J. Mitra, X. Huang, and S. M. Hu:"Global contrast based salient region detection", CVPR (2011)

(5) J. Harel, C. Koch, and P. Perona. Graph-based visual saliency:"Graph-based visual saliency", NIPS (2006)

(6) X. Hou and L. Zhang:"Saliency detection: A spectral residual approach", CVPR (2007)

(7) T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. Shum:"Larning to detect a salient object", IEEE PAMI (2011)

(8) J. Yang and M. Yang:"Top-down visual saliency via joint CRF and dictionary learning", CVPR (2011)

(9) C.Yang, L. H. Zhang, H. C. Lu, X. Ruan, M. M. Yang:"Saliency Detection via Graph-Based Manifold Ranking", CVPR (2013)

(10) J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba:"SUN Database: Large-scale Scene Recognition from Abbey to Zoo", ICCV (2010)

(11) A. Borji, D. N. Sihite, L. Itti:"Salient Object Detection: A Benchmark", ECCV (2012)

(12) R. Subramanian, H. Katti, N. Sebe, M. Kankanhalli, T. S. Chua:"An eye fixation database for saliency detection in images", ECCV (2010)

(13) N. D. B. Bruce, J. K. Tsotsos:"Saliency, attention, and visual search: An information theoretic approach", Journal of vision (2009)

(14) G. Kootstra, A. Nederveen, D. B. Bart:"Paying attention to symmetry", BMVC (2008)

(15) T. Judd, K. Ehinger, F. Durand, and A. Torralba:"Learning to predict where humans look", ICCV (2009)

(16) H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng:"Automatic salient object segmentation based on context and shape prior", BMVC (2011)

(17) S. Goferman, L. Zelnik-Manor, and A. Tal:"Context-aware saliency detection", CVPR (2010)

(18) Y. Xie, H. C. Lu, and M. M. Yang:"Bayesian saliency via low and mid level cues", TIP (2013)

(19) Y. Xie and H. C. Lu, and M. M. Yang:"Visual saliency detection based on Bayesian model" ICIP (2011)

(20) E. Rahtu, J. Kannala, M. Salo, and J. Heikkil:"Segmenting salient objects from images and videos" ECCV (2010)

(21) X. Shen and Y. Wu:"A unified approach to salient object detection via low rank matrix recovery" ICCV (2012)

(22) L. Itti, C. Koch, and E. Niebur:"A model of saliency-based visual attention for rapid scene analysis" PAMI (1998)

(23) Y. Zhai and M. Shah:"Visual attention detection in video sequences using spatiotemporal cues" ICME (2006)

Xiang Ruan received the B.E. degree from Shanghai Jiao Tong University, Shanghai, China, in 1997, and the M.E and Ph.D. degrees from Osaka City University, Osaka, Japan, in 2001 and 2004, respectively. He is currently a Research Engineer with OMRON Corporation, Kyoto, Japan. His current research interests include computer vision, machine learning, and image processing.

Huchuan Lu received the M.Sc. degree in signal and information processing and the Ph.D. degree in system engineering, Dalian University of Technology (DUT), Dalian, China, in 1998 and 2008, respectively. Since 1998, he has been a faculty of the same university. Since 2011, he has been the professor in the School of Information and Communication Engineering of DUT. His current research interests include the areas of computer vision and pattern recognition. In recent years, he focuses on visual tracking and segmentation. Prof. Lu is a member of the ACM and an associate editor of the IEEE T-SMC Part:B.

Na Tong is a second year Master student, with School of Information and Communication Engineering, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, 116024, P.R. China.