A DATASET AND EVALUATION METHODOLOGY FOR VISUAL SALIENCY IN VIDEO

Jia Li[1,2] *Yonghong Tian*[3] *Tiejun Huang*[3] *Wen Gao*[1,3]

[1]Key Lab of Intell. Info. Process, Inst. of Comput. Tech., Chinese Academy of Sciences, China
[2]Graduate University of Chinese Academy of Sciences, China
[3]Institute of Digital Media, School of EE & CS, Peking University, China

ABSTRACT

Recently, visual saliency has drawn great research interest in the field of computer vision and multimedia. Various approaches aiming at calculating visual saliency have been proposed. To evaluate these approaches, several datasets have been presented for visual saliency in images. However, there are few datasets to capture *spatiotemporal* visual saliency in video. Intuitively, visual saliency in video is strongly affected by temporal context and might vary significantly even in visually similar frames. In this paper, we present an extensive dataset with 7.5-hour videos to capture *spatiotemporal* visual saliency. The salient regions in frames sequentially sampled from these videos are manually labeled by 23 subjects and then averaged to generate the ground-truth saliency maps. We also present three metrics to evaluate competing approaches. Several typical algorithms were evaluated on the dataset. The experimental results show that this dataset is very suitable for evaluating visual saliency. We also discover some interesting findings that would be addressed in future research. Currently, the dataset is freely available online together with the source code for evaluation.

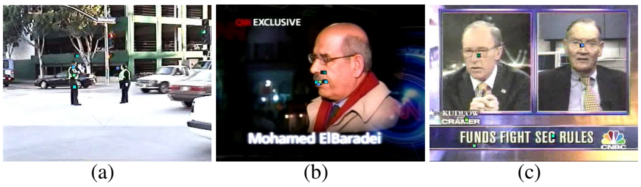
Index Terms— Visual saliency, dataset, evaluation metrics, saliency map, salient regions

1. INTRODUCTION

Visual saliency is the distinct subjective perceptual quality which makes some items in the scene stand out from their neighbors and immediately grab our attention [1]. Recently, visual saliency has drawn great research interest in the field of computer vision and multimedia. It is often assumed that visual saliency is extracted and represented in an explicit saliency map, which serves to determine the salient regions. Such salient regions can be selected to assist applications such as content-relevant advertising [2] and shot matching [3].

With the numerous algorithms proposed to estimate visual saliency, evaluation becomes a critical issue. An evaluation strategy for visual saliency involves at least two aspects: appropriate criteria and metrics, and benchmark datasets. Among them, benchmark datasets can provide us an open testing platform to choose from many different proposals and to evaluate new approaches against older ones. Usually, such benchmark dataset for visual saliency could be constructed through subjective evaluation. For example, Liu *et al.* [4] had 9 subjects to manually label the most salient object in each image. Regional saliency is obtained by averaging the labeling results. Bruce *et al.* [5] recorded the eye tracking data of 20 subjects when they were watching the 120 natural images, each for 4 seconds. Then the fixation density maps were generated. However, the two approaches might encounter difficulties for *spatiotemporal* visual saliency in video, which is strongly affected by temporal context and might vary significantly even in visually similar frames. It was often inaccurate to directly label the salient objects by treating each frame as an independent image. Itti *et al.* [6] constructed a video dataset by recording eye tracking data of 8 subjects. However, as shown in Fig.1 (a) and (c), human beings tend to attend multiple objects synchronously when watching videos. It is difficult to recover *regional* saliencies of multiple objects merely from several eye fixations recorded in the short interval of each frame (See Fig.1 (b) and (c)).

In this paper, our goal is to construct an extensive dataset on visual saliency in video. We focus on revealing the *regional* saliency in sequential frames. Firstly, we collected a dataset consisting of 431 short videos with a total length of 7.5 hours. According to psychological evidences in [7] that "interesting" objects, rather than early features, guide human attention, the salient objects in sampled frames are manually labeled by 23 subjects. The labeling results are then averaged to generate the ground-truth saliency maps (GSMs). Three metrics are presented to evaluate visual saliency estimation approaches on the dataset. Experiments performed with several typical methods show that this dataset is suitable for evaluating visual saliency. We also discover some interesting findings that would be addressed in future research.



(a) (b) (c)

Fig. 1. Examples for effects of eye tracker in video [6].

The rest of the paper is organized as follows. Sec. 2 briefly reviews the evaluation methodologies in related work. In Sec. 3, the collection and labeling of the dataset are presented. Sec. 4 proposes some evaluation metrics. Sec. 5 shows experiments of several typical algorithms on the dataset and Sec. 6 concludes the paper.

2. RELATED WORK

This section is devoted to reviewing the evaluation methods on visual saliency. In general, the evaluation methodologies can be divided into three categories: point-based evaluation (e.g., [1], [5], [8, 9]), region-based evaluation (e.g., [4], [10, 11]), and subjective evaluation (e.g., [3], [12]).

Traditionally, visual saliency calculation was modeled as a gaze prediction problem. In Itti's early work [1], gaze points were selected from the estimated saliency maps (ESMs) to check whether they could locate the most salient object in images. However, Bruce *et al.* [5] argued that it might be not appropriate to select discrete fixations from the ESMs. Instead, they evaluated the ESMs with the recorded eye fixations in images by computing the receiver operator characteristic (ROC) scores. Similarly, Kienzle *et al.* [8] computed the ROC scores with respect to the eye fixations in videos. In [9], Itti *et al.* sampled small patches from the ESMs around the recorded fixations and random points respectively, and quantified the difference between their distributions using the *Kullback-Leibler distance* (KLD). In general, these evaluation approaches are effective for evaluating predicted gaze but not for the *regional* saliency.

Often, *regional* saliency was evaluated by comparing the estimated and labeled salient regions. In [4, 10, 11], the masks of the salient objects in natural images were firstly generated by combing the labeling results of multiple subjects. The mask maps were then used for regional saliency evaluation. Unfortunately, there are no datasets and evaluation methods for regional saliency evaluation in video.

Subjective evaluation was also frequently used. In [3] and [12], for example, subjective scores on the estimated saliencies were reported on three levels: "Good", "Acceptable" and "Failed". Clearly, subjective evaluation approach can not be extended to large scale datasets.

In summary, although salient region detection is an important issue in many multimedia tasks such as scalable video coding, content analysis, and content-relevant advertising, there are few video datasets and evaluation methods on *spatiotemporal* visual saliency in video.

3. DATASET DESIGN

3.1. Data Collection

The main goal of data collection is to cover videos of various scenes. The dataset is collected as follows:

1) ***Surveillance video***. In most cases, surveillance video contains static backgrounds and dynamic salient objects, which can be used for visual saliency analysis. In our dataset, surveillance videos are selected from the CAVIAR dataset (http://homepages.inf.ed.ac.uk/rbf/CAVIAR/).

2) ***Artificial video***. To explore the differences of visual saliency in natural and artificial scenes, we select artificial video clips from 2D and 3D cartoons.

3) ***Natural video with artificial parts***. Usually, the artificial parts in natural video such as captions and logos have a strong impact on visual saliency. We collect videos with artificial parts from TRECVID 2006/2007 and the Internet.

4) ***Natural video***. Similarly, we select natural videos with no artificial parts such as overlaid captions and logos from TRECVID 2006/2007 and the Internet.

In total, our dataset contains 431 short videos with a total length of 7.5 hours. In total, 764,806 frames are involved. The dataset mainly covers videos from six genres: *documentary*, *ad*, *cartoon*, *news*, *movie* and *surveillance*.

3.2. Data Labeling

Among the collected videos, the salient (foreground) objects in *surveillance* videos have already been labeled. For other videos, we assign 23 subjects to label them (17 men, 6 women, aged between 21 and 37 years, 10-23 subjects for each clip). Since it is invincible to manually generate the GSM for each frame, we sample 62,356 key frames from the selected videos (I frames for MPEG videos or sampling one frame out of every 15 frames for other videos).

In the labeling process, subjects are first instructed to watch a short video. Then the key frames of the former video are displayed again in chronological order. Subjects are asked to label all regions that they thought to be "salient" in previous watching with one or multiple rectangles. Labeling differences might occur at borders, which can be greatly reduced by combining the labels of multiple subjects. For example, we select 10 subjects and divide them into two groups A and B. The labeling difference between the two groups is computed as:

$$\mathrm{LD}(A,B) = \sum |N_A - N_B| / \sum (N_A + N_B), \qquad (1)$$

where N_A and N_B denote the numbers of subjects in A and B who have labeled the same position. On average the difference between two subjects is 39.68%± 5.22%, whereas that between two groups is reduced to 23.05% ± 1.77%. It shows that the labeling differences between groups are much smaller. After labeling, the GSMs can be obtained by combining the labeling results of all subjects. For a frame F_i, $a_{i,k}$ subjects have selected the k-th macro block as "salient". Thus, the ground-truth visual saliency of the k-th macro block can be computed as:

$$g_{i,k} = a_{i,k} / \sum_j a_{i,j}. \qquad (2)$$

To smooth the edges of the labeling rectangles, the GSMs generated from (2) are convolved with a 2D Gaussian kernel ($\sigma = 8$). Some examples of GSMs are shown in Fig. 2. From Fig.2 (a-c), we see that our labeling method can reveal the saliencies of multiple regions. Moreover, the effect of temporal context is taken into account in the labeling. For example, we can see from Fig.2 (d-f) that the GSMs of the three consequent frames are changing even they are visually similar. Meanwhile, objects (e.g., dog, gate) with rare inter-frame variation are almost ignored.

3.3. Advantages and Drawbacks of Our Dataset

Our dataset provides a feasible benchmark for evaluating the *spatiotemporal* saliency in video. The effects of temporal context can be effectively reflected in the GSMs, since subjects are asked to label the salient objects by recalling what they have seen in the previous video watching process instead of directly labeling each frame. Moreover, it is the salient objects that are labeled instead of interesting points. Thus the dataset is also useful for evaluating saliency-based applications such as content-relevant advertising.

One drawback of our dataset owes to the rectangular labeling. To solve this problem, we will incorporate image segmentation approaches to improve the borders of labeled objects in the future. Another drawback is that it is inefficient to evaluate the prediction algorithms of gazes and interesting points since we only provide *regional* saliency.

4. EVALUATION METHODOLOGY

Given the dataset, we need a set of appropriate criteria and metrics to quantitatively evaluate the performances of different approaches for visual saliency estimation. Intuitively, the evaluation metrics should take the following criteria into account. 1) *Consistency*: the saliency distributions in the GSM and the ESM should be perfectly consistent; 2) *Regional similarity*: the salient regions extracted from the GSM and the ESM should precisely overlap; and 3) *Compactness*: the saliencies of the ESM should converge into limited regions. Consequently, we present the following three metrics:

1) ***Precision***. Suppose that the GSM and ESM of the i-th frame are expressed as normalized vectors $\mathbf{g}_i$ and $\mathbf{g}'_i$, we can evaluate their consistency (precision) with cross entropy:

$$P_i = \exp\left[-\frac{1}{2}\sum_k (g_{i,k}\log^{\frac{2g_{i,k}}{g_{i,k}+g'_{i,k}}} + g'_{i,k}\log^{\frac{2g'_{i,k}}{g_{i,k}+g'_{i,k}}})\right], \tag{3}$$

where $\exp(\cdot)$ function is used to normalize the precision into [0, 1]. We use the *Jensen-Shannon Divergence* (JSD) here to evaluate the cross entropy instead of KLD since JSD is symmetric and would work better in the cases of small or zero possibilities. The precision score would be larger if the ESM better approximates the GSM.

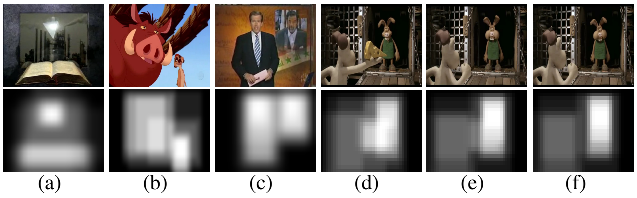


Fig. 2. Some representative examples of frames and GSMs.

2) ***ROC***. ROC is effective to evaluate a binary classifier system. Here we use it to evaluate the similarity of ESMs and GSMs. In the evaluation process, both the ESM and the GSM are first quantized with different thresholds to obtain the salient regions. The thresholds can be randomly selected respectively for the ESM and the GSM. After that, the ROC curve is plotted as the *false positive rate* vs. *true positive rate*, and the ROC score is defined as the area under the ROC curve. Perfect prediction corresponds to a score of 1.

3) ***Compactness***. By measuring the entropy of the ESM, the compactness can be computed as:

$$C_i = \exp\left(-\sum_k g'_{i,k}\log^{g'_{i,k}} / \sum_k g_{i,k}\log^{g_{i,k}}\right), \tag{4}$$

where the entropy of the GSM is used as the reference. Similarly, we use $\exp(\cdot)$ function to normalize the compactness into [0, 1]. According to (4), perfect estimation should generate several salient regions while suppressing other regions. These estimation should have a large compactness close to 1.

5. EXPERIMENTAL RESULTS

We also conducted experiments to verify whether our dataset and evaluation metrics are effective. Six algorithms [1, 9, 10, 11, 12, 13] were performed in the experiments. Among them, Itti'98 [1] and Itti'01 [13] are two well-known approaches which use the multi-scale center-surround contrasts for saliency estimation; Itti'05 [9] and Zhai'06 [12] focus mainly on finding irregularities in the inter-frame variations; Hou'07 [10] and Guo'08 [11] explore the amplitude or phase spectrum to select the irregular stimulus as "salient". The dataset and C++ source code for evaluation are available at http://www.jdl.ac.cn/user/jiali/RSD_Dataset.htm.

The overall results using the evaluation metrics proposed in Sec. 4 are listed in Table 1. Some representative examples are given in Fig.3. Since both Itti'98 [1] and Itti'01 [13] only "pop-out" the most salient points in the saliency map using "winner-take-all" strategy, it is meaningless to compare these salient points with ground-truth regions. As a result, the ROC scores are not given for the two algorithms.

From Table 1 and Fig.3, we can see that all the six algorithms perform well in some scenes. For example, Itti'05 [9] performs well in scenes with static background (e.g., the first, second and fifth columns in Fig.3 (d)), while Zhai' 06 [12] can effectively detect the large objects with rigid motion (e.g., the first, second and fourth columns in Fig.3 (g)).

This indicates that our dataset are suitable for evaluating visual saliency. However, none of the algorithms performs perfectly on the whole dataset since our dataset covers a variety of complex scenes. It is challenging to establish a unified framework for visual saliency estimation on our dataset. By analyzing the failures of these algorithms, we discover several interesting issues that would be addressed in the future:

1) ***Noise suppression***. A good visual saliency estimation algorithm should effectively suppress noises while preserving salient regions. As shown in Fig.3 (e) and (f), the results of Hou'07 [10] and Guo'08 [11] have many noises because they tend to find all irregular "edges" in amplitude or phase spectrums. This is the reason why they have low compactness scores but high ROC and precision scores. In contrast, Itti'98 [1] and Itti'01 [13] over-suppress the noises using the "winner-take-all" strategy, leading to high compactness scores and low precision scores.

2) ***Generalization***. A good algorithm should perform well in various scenes. As shown in the third column in Fig.3 (d), for example, Itti'05 [9] would fail in scenes with static salient objects since it tends to detect strong inter-frame variations. Similarly, Zhai'06 [12] can hardly detect small salient objects, as shown in the fifth column in Fig.3 (g).

3) ***Saliency estimation of internal homogenous regions***. Beyond detecting the irregularities in images and videos, the homogeneous regions inside the salient objects should also be assigned with high saliencies.

6. CONCLUSION

In this paper, we present an extensive video dataset and an evaluation methodology for visual saliency in video. The experiment results show that our dataset is effective and especially suitable for evaluating regional saliency in video. We have also discover several interesting issues that would be addressed in the future. One direction for improving this dataset is to replace the rectangular labeling with labeling object borders. This could be done under the assistance of image segmentation approaches.

7. ACKNOWLEDGEMENT

The work is supported by grants from Chinese NSF under contract No. 60605020 and No. 90820003, National Hi-Tech R&D Program (863) of China under contract 2006AA010105, and National Basic Research Program of China under contract No. 2009CB320906.

8. REFERENCES

[1] L. Itti, C.Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. PAMI*, 20(11): 1254–1259, 1998.

[2] T. Mei, X.-S. Hua, and S. Li, "Contextual In-Image Advertising," In *ACM MM*, 2008.

Table 1. Experiment results of different algorithms

Algorithm	Precision	ROC	Compactness
Itti'98 [1]	0.458	-	0.536
Itti'05 [9]	0.655	0.596	0.3834
Hou'07 [10]	0.728	**0.661**	0.352
Guo'08 [11]	**0.775**	0.645	0.320
Zhai'06 [12]	0.682	0.610	0.401
Itti'01 [13]	0.449	-	**0.613**

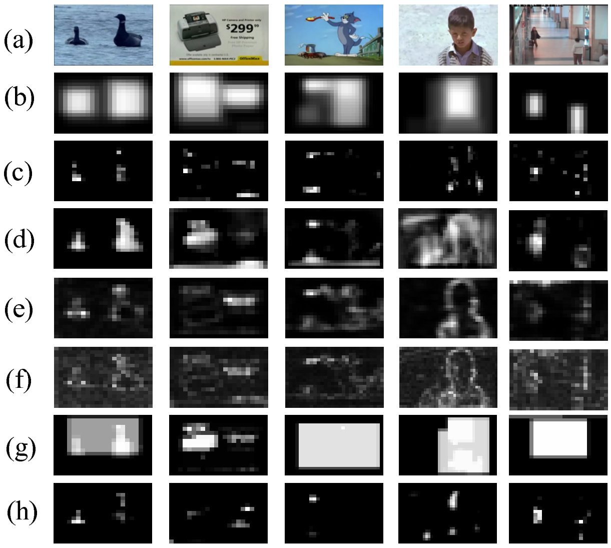


Fig. 3. Typical examples of visual saliency prediction. (a) original frames; (b) GSMs; (c)-(h) results of algorithms [1],[9],[10],[11],[12] and [13].

[3] S. Li and M.-C. Lee, "Efficient Spatiotemporal Attention-driven Shot Matching," In *ACM MM*, 2007.

[4] T. Liu, J. Sun, N.-N. Zheng, X. Tang and H.-Y. Shum, "Learning to Detect a Salient Object," In *CVPR*, 2007.

[5] N. Bruce, and J.K. Tsotsos, "Saliency Based on Information Maximization," *Advances in Neural Information Processing Systems*, 18: 155-162, 2006.

[6] L. Itti, "CRCNS Data Sharing: Eye Movements During Free-viewing of Natural Videos," In *Collaborative Research in Computational Neuroscience Annual Meeting*, 2008.

[7] Wolfgang Einhauser, M. Spain and P. Perona, "Objects Predict Fixations Better Than Early Saliency," *J. Vision*, 8(14):18, 1-26, 2008.

[8] W. Kienzle, B. Scholkopf, F. Wichmann and M. O. Franz, "How to Find Interesting Locations in Video: a Spatiotemporal Interest Point Detector Learned From Human Eye Movement," In *DAGM*, 2007.

[9] L. Itti and P. Baldi, "A Principled Approach to Detecting Surprising Events in Video," In *CVPR*, 2005.

[10] X. Hou and L. Zhang, "Saliency Detection: A spectral Residual Approach," In *CVPR*, 2007.

[11] C.L. Guo, Q. Ma and L.M. Zhang, "Spatio-temporal Saliency Detection using Phase Spectrum of Quaternion Fourier Transform," In *CVPR*, 2008.

[12] Y. Zhai and M. Shah, "Visual Attention Detection in Video Sequences Using Spatiotemporal Cues," In *ACM MM*, 2006.

[13] L. Itti and C.Koch, "Computational Modeling of Visual Attention," *Nature Review Neuroscience*, 2(3):194–203, 2001.