Why You Trust in Visual Saliency

Edoardo Ardizzone, Alessandro Bruno(✉), Luca Greco, and Marco La Cascia

DICGIM, Università degli Studi di Palermo, Viale delle Scienze bd.6 90128, Palermo, Italy
{edoardo.ardizzone,alessandro.bruno15,luca.greco,
marco.lacascia}@unipa.it

Abstract. Image understanding is a simple task for a human observer. Visual attention is automatically pointed to interesting regions by a natural objective stimulus in a first step and by prior knowledge in a second step. Saliency maps try to simulate human response and use actual eye-movements measurements as ground truth. An interesting question is: how much corruption in a digital image can affect saliency detection respect to the original image? One of the contributions of this work is to compare the performances of standard approaches with respect to different type of image corruptions and different threshold values on saliency maps. If the corruption can be estimated and/or the threshold is fixed, the results of this work can also be used to help in the selection of a method with best performance.

Keywords: Saliency maps · Image corruption · Image compression

1 Introduction

The problems of automatic categorization and understanding of digital images is still open even tough for a human observer are quite simple problems to solve. Humans use both purely visual features (pre-attentive) and prior knowledge on the world to assign a sense to a picture.

The term saliency refers to visual characteristic of interest for a human observer. The aim of visual saliency detection methods is to build a saliency map that tries to replicate the human visual system (HVS) behavior in the visual attention process. Salient parts of a scene are those regions that create a strong visual response and polarize attention. Human attention is the sum of factors coming from two different stimuli: the first one depends exclusively on the characteristics of the image, the second one is subjective for the observer and is related to his will (it is task-dependent). The objective stimulus (pre-attentive) is excited by the physical characteristics such as brightness, color, shape and has a bottom-up activation. In many situations, however, the largest contribution is given by the top-down process, because the focus of the attention is largely influenced by the knowledge obtained by learning the probabilistic structure of the scene.

Methods of the state of the art observe the actual behavior of human eye tracking its movements with special glasses (eye-tracker) and use observed movements as benchmark for proposed approaches. In most cases, the image quality is not mentioned in saliency map generation.

V. Murino et al. (Eds.): ICIAP 2015 Workshops, LNCS 9281, pp. 589–596, 2015.
DOI: 10.1007/978-3-319-23222-5_71

In this paper we want to investigate the robustness of some popular visual saliency methods against image degradation, such as jpeg compression and noise. In the next sections of the paper we try to answer this question by analyzing the performances of some saliency map generation algorithms with respect several kind of image corruptions. A similar work is [10], that focus on the performances of several saliency models on a corrupted database [11] for the quality assessment of digital images. Authors performed the analysis of the ROC and other metrics using five different saliency extraction methods in blurred, compressed (JPEG) and noisy images. In this case the benchmark dataset is formed by the results of eye-tracking both in original than in distorted images. The remainder of the paper is organized as follows: Section 2 shows related works in this field both in saliency extraction than in methods comparison; Section 3 describes which metrics can be used in saliency performance evaluation; Section 4 gives the result of different methods on original and corrupted images. Section 5 contains conclusions.

2 Visual Saliency Estimation

Models for visual saliency detection and extraction are inspired by human visual system and tend to reproduce the dynamic modifications of cortical connectivity for scene perception. Generally Saliency approaches can be divided in three main groups: Bottom-up, Top-down, Hybrid.

In Bottom-up methods, human attention is considered a cognitive process that selects most unusual part (i.e. distinct objects, contours) of an environment while ignoring most common aspects (i.e. uniform background).

A fundamental bottom-up and stimulus driven approach, proposed by [1], for Saliency detection adopted multi-scale analysis of the image. Multi-scale image features are combined into a single topographical saliency map. A dynamical neural network selects attended locations in order of decreasing saliency.

Harel [2] saliency method is based on a biologically plausible graph based model, consists of two steps: activation maps on certain feature channels and normalization which highlights conspicuity. The method proposed in [3] is based on parallel extraction of different feature maps using center-surround differences.

In Top-down approaches [4,5] the visual attention process is considered task dependent, and the observer's expectations and wills analyzing the scene are the reason why a point is fixed rather than others.

Generally Hybrid systems for saliency use the combination of bottom-up and top-down stimulus. In many hybrid approaches [6,7] Top-down layer is used to refine the noisy map extracted from Bottom-up layer. For example the top-down component in [6] is face detection. Chen et al. [7] used a combination of face and text detection and they found the optimal solutions through branch and bound technique.

A state of the art well known hybrid approach was proposed by Judd et al. [8] in addition to database [9] of eye tracking data from 15 viewers. Low, middle and high-level features of this data have been used to learn a model of saliency.

A comparative study that evaluates the performances of 13 state-of-the-art saliency maps has been reported in [12]. Test images are composed of a target object and a cluttered background and the conspicuity of the image is assumed to rely in the target area. A new metric is also proposed and compared with previous models.

The work [13] is a short survey on current state of the art. It contains some formal definitions on three different type of approaches (bottom-up, top-down, hybrid) and an overview on existing methods. Then, authors offer a description of publicly available datasets and the performance metrics used. Finally there is a description of the computational methods used in previous described literature of saliency extraction.

3 Metrics

Benchmarks in saliency extraction are formed by a list of fixation points and a continue map representing how much a pixel is salient. Generally, this map is obtained convolving a gaussian filter across the real points. On the other hand, saliency extraction methods give as result a map showing the value of the saliency for each pixel of the image. Using the continue ground truth as the reference one permits only an approximate comparison. A deeper investigation include the analysis using several thresholds on the map, so appreciating the similarity of the results centering on blobs corresponding to real fixation points. In experimental results section a deeper analysis about the robustness of saliency extraction methods against image degradation is given. We performed several experiments using different approaches of saliency computation.

3.1 Considered Metrics

There are many ways to compare two, normalized, saliency maps. Assuming that a pixel shows "positive" result if his value is over a threshold and a "negative" result otherwise, the classical information retrieval can be used.

Most known metrics are:

- Precision (P): the ratio of true positives (TP) and the sum of true positives and false positives (FP). It measure how much the pixels considered as salient in the computed map are salient also in the ground truth.

$$P = \frac{(TP)}{(TP+FP)} \tag{1}$$

- Recall (R):the ratio of true positives and the sum of true positives and false negatives (FN). It measure how much the regions considered as salient in the ground truth are present in the compared map.

$$R = \frac{(TP)}{(TP+FN)} \tag{2}$$

- Detection Measure (DM): the product of P and R. It prevents incorrect cases of high precision (i.e. a few of small blobs) with a low recall and the inverse situation of a high recall (i.e. considering salient the entire image)

$$DM = P \times R \tag{3}$$

- F-Measure (F): the harmonic mean of P and R. It uses a different approach in limiting the problems also addressed by DM.

$$F = 2 \times \frac{(P \times R)}{(P + R)} \tag{4}$$

4 Results

The contribution of this work is mostly a comparative evaluation of three well-known method ITTI [1], GBVS [2], and TORRALBA [8]. The attention is focused on the results obtained using a corrupted version of the test images. We used the dataset provided in [8], where benchmark fixation maps are created tracking the real movements of the eye. In our test we used these maps to see how different approaches can simulate the real response of the eye with or without corruption in the images. In particular, we considered the effect of compression and two common type of noise.

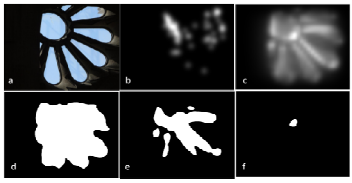

Fig. 1. Original Image (a), Ground truth (b), Saliency map (c), Resulting maps with thresholds 0.1 (d), 0.5 (e) and 0.75 (f)

4.1 Saliency Map Accuracy

A first evaluation of the methods is done analyzing how much resulting maps are similar to real fixation maps. Continue maps of the benchmark are generated from fixation maps by convolving gaussian filters on the fixation points, so approximating the result for a unknown observer. The perfect fitting is not very important in continue maps, a best way to compare maps is looking at the regions of concentration of saliency and seeing how much these regions are corresponding. Assuming a saliency normalized map in range 0-1 this can be done thresholding the map and obtaining a binary version, for example assigning the value 0 to original values under the threshold and 1 otherwise. An example of this is shown in Fig. 1.

First comparison is done calculating P,R,DM and FM for the methods using 4 values of threshold: 0.1, 0.25, 0.5, 0.75. Results are shown in Fig. 2.

Observing this values, it is possible to notice that Bottom-up methods (GBVS, ITTI) show better performances with lower level of thresholding (0.1 to 0.50) in P, DM and FM metrics and TORRALBA exceed others only in R. On the other hand, TORRALBA method normally gives a more sparse map (covering a large part of the image) so high values of recall is an expected result. TORRALBA method outperforms ITTI and GVBS with high values of thresholding. Its resulting map, using various low- mid- and high- level features, covers a large part of the image but contains high concentrated saliency values in blobs close to the actual fixation points.

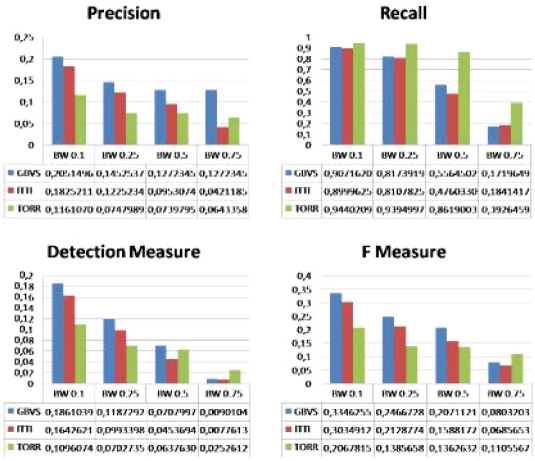

Fig. 2. Graphs of P, R, DM and FM for different thresholds

4.2 Effect of Noise

In this section we study the effect of additive noise in saliency calculation. The saliency map of the original image of [8] is compared with the one extracted from eight noisy versions: four with Gaussian noise and four with salt and pepper noise having variance 0.01, 0.1, 0,5 and 1 respectively. Larger is the variance, stronger is the corruption of the image.

In Table 1 P, R and FM values are shown for GBVS, ITTI and TORRALBA methods. The results show that R of all methods increases with noise because the resulting maps are more sparse and cover a larger part of the image. Results for TORRALBA are the best ones. On the other hand, P has better results with bottom-up approach for low thresholds (and best performer GBVS) but TORRALBA is preferable for higher values and is generally more stable with noise, leading to a P value similar to the original in noisy images. GBVS shows the best result for low thresholds and always outperforms ITTI;TORRALBA is less affected by noise, giving best results only for high thresholding.

Table 1. Performances of GBVS, ITTI and TORRALBA with noise introduction

GBVS	Precision				Recall				F-measure			
	TH 01	TH 0.25	TH 0.5	TH 0.75	TH 01	TH 0.25	TH 0.5	TH 0.75	TH 01	TH 0.25	TH 0.5	TH 0.75
original	0,176	0,123	0,087	0,051	0,976	0,928	0,582	0,236	0,287	0,206	0,140	0,074
gauss 0.01	0,174	0,121	0,086	0,050	0,976	0,926	0,583	0,240	0,275	0,191	0,133	0,075
gauss 0.1	0,165	0,111	0,079	0,048	0,979	0,947	0,606	0,266	0,251	0,162	0,114	0,066
gauss 0.5	0,152	0,097	0,071	0,045	0,983	0,959	0,635	0,311	0,259	0,171	0,122	0,074
gauss 1	0,147	0,090	0,065	0,039	0,986	0,965	0,647	0,318	0,790	0,209	0,147	0,075
S&P 0.01	0,175	0,122	0,087	0,051	0,976	0,928	0,581	0,233	0,289	0,208	0,142	0,076
S&P 0.1	0,169	0,115	0,083	0,051	0,977	0,935	0,589	0,243	0,281	0,198	0,137	0,076
S&P 0.5	0,153	0,098	0,069	0,039	0,983	0,955	0,605	0,254	0,226	0,131	0,080	0,030
S&P 1	0,130	0,071	0,044	0,016	0,991	0,980	0,643	0,306	0,259	0,173	0,118	0,062

ITTI	Precision				Recall				F-measure			
	TH 01	TH 0.25	TH 0.5	TH 0.75	TH 01	TH 0.25	TH 0.5	TH 0.75	TH 01	TH 0.25	TH 0.5	TH 0.75
original	0,156	0,110	0,074	0,038	0,978	0,872	0,520	0,168	0,248	0,183	0,243	0,059
gauss 0.01	0,146	0,106	0,072	0,038	0,984	0,891	0,530	0,168	0,216	0,148	0,201	0,053
gauss 0.1	0,123	0,083	0,057	0,032	0,996	0,948	0,568	0,187	0,194	0,102	0,126	0,024
gauss 0.5	0,111	0,060	0,041	0,022	1,000	0,980	0,585	0,190	0,196	0,111	0,149	0,038
gauss 1	0,109	0,055	0,034	0,013	1,000	0,994	0,581	0,168	0,262	0,187	0,247	0,059
S&P 0.01	0,147	0,103	0,071	0,038	0,979	0,875	0,570	0,166	0,250	0,178	0,738	0,060
S&P 0.1	0,133	0,088	0,061	0,034	0,990	0,903	0,533	0,160	0,230	0,156	0,209	0,053
S&P 0.5	0,111	0,067	0,038	0,014	1,000	0,985	0,567	0,140	0,191	0,095	0,108	0,013
S&P 1	0,107	0,050	0,028	0,007	1,000	0,999	0,668	0,336	0,197	0,115	0,139	0,024

TORRALBA	Precision				Recall				F-measure			
	TH 01	TH 0.25	TH 0.5	TH 0.75	TH 01	TH 0.25	TH 0.5	TH 0.75	TH 01	TH 0.25	TH 0.5	TH 0.75
original	0,109	0,059	0,060	0,062	1,000	0,998	0,744	0,489	0,193	0,108	0,102	0,095
gauss 0.01	0,109	0,058	0,060	0,061	1,000	0,999	0,745	0,491	0,193	0,106	0,098	0,090
gauss 0.1	0,108	0,057	0,057	0,058	1,000	0,998	0,743	0,487	0,192	0,102	0,091	0,081
gauss 0.5	0,108	0,055	0,053	0,050	1,000	0,999	0,748	0,497	0,192	0,103	0,094	0,085
gauss 1	0,108	0,054	0,051	0,047	1,000	0,998	0,762	0,525	0,194	0,110	0,103	0,095
S&P 0.01	0,109	0,058	0,060	0,062	1,000	0,998	0,741	0,484	0,194	0,109	0,102	0,096
S&P 0.1	0,108	0,057	0,058	0,060	1,000	0,999	0,738	0,478	0,193	0,107	0,099	0,090
S&P 0.5	0,108	0,055	0,054	0,054	1,000	0,998	0,746	0,493	0,191	0,095	0,077	0,050
S&P 1	0,107	0,050	0,039	0,077	1,000	0,998	0,748	0,483	0,192	0,103	0,096	0,089

4.3 Effect of Compression

In this section are presented the performances using compressed (JPEG) images. The maps are calculated using the compressed version of the original image with different quality factors. Results of P for GBVS and TORRALBA methods are shown in following figures. Results for ITTI are not shown because similar to GVBS.The visual effects of compression depend on morphological surface of the images, for this reason we analyzed several effects on images with very different morphological surfaces. The testset we used consists of a lot of images with very different background and foreground compositions: a simple object in the foreground and a homogenous background, many object in the foreground and a chaotic background.

The results show that, in terms of P and R, saliency methods are robust against different compression rates. In practical terms this means that substantially the jpeg compression do not affect the performance of saliency detection methods. In greater details, GBVS approach overcomes the others method Precision results, while

TORRALBA shows the lower Precision values. On the other side, TORRALBA shows the higher values of Recall (because, generally the resulting saliency maps cover larger regions of the image) that, however, are close to GBVS and ITTI ones.

Furthermore, observing preliminary experimental results, we notice that, for every methods, Precision values of saliency maps with several thresholds are really close to each other.

Table 2. Performances of GBVS and TORRALBA with compressed images

GBVS	Precision				Recall				F-measure			
	TH 01	TH 0.25	TH 0.5	TH 0.75	TH 01	TH 0.25	TH 0.5	TH 0.75	TH 01	TH 0.25	TH 0.5	TH 0.75
JPEG 1%	0,186	0,137	0,099	0,052	0,959	0,846	0,517	0,180	0,298	0,717	0,147	0,071
JPEG 5%	0,182	0,133	0,100	0,061	0,958	0,879	0,551	0,206	0,294	0,212	0,147	0,083
JPEG 10%	0,185	0,134	0,107	0,052	0,958	0,880	0,590	0,188	0,295	0,717	0,157	0,073
JPEG 20%	0,186	0,137	0,054	0,049	0,959	0,872	0,904	0,181	0,297	0,215	0,094	0,068
JPEG 40%	0,187	0,138	0,110	0,050	0,959	0,872	0,588	0,186	0,298	0,217	0,160	0,070
JPEG 60%	0,186	0,136	0,109	0,050	0,958	0,870	0,593	0,185	0,297	0,215	0,159	0,069
JPEG 80%	0,187	0,137	0,110	0,050	0,958	0,869	0,589	0,184	0,298	0,216	0,160	0,069
original	0,187	0,138	0,110	0,050	0,958	0,869	0,589	0,185	0,298	0,216	0,160	0,069
TORRALBA	**Precision**				**Recall**				**F-measure**			
	TH 01	TH 0.25	TH 0.5	TH 0.75	TH 01	TH 0.25	TH 0.5	TH 0.75	TH 01	TH 0.25	TH 0.5	TH 0.75
JPEG 1%	0,111	0,058	0,052	0,050	1,000	0,998	0,916	0,370	0,196	0,108	0,095	0,081
JPEG 5%	0,111	0,058	0,053	0,055	1,000	0,998	0,920	0,402	0,196	0,109	0,095	0,086
JPEG 10%	0,111	0,058	0,050	0,060	1,000	0,998	0,928	0,457	0,196	0,108	0,090	0,096
JPEG 20%	0,111	0,059	0,021	0,068	1,000	0,997	1,000	0,453	0,197	0,109	0,041	0,103
JPEG 40%	0,111	0,059	0,051	0,064	1,000	0,998	0,925	0,446	0,197	0,109	0,091	0,099
JPEG 60%	0,111	0,059	0,051	0,064	1,000	0,997	0,926	0,449	0,197	0,109	0,091	0,101
JPEG 80%	0,111	0,059	0,051	0,066	1,000	0,998	0,924	0,441	0,197	0,109	0,092	0,102
original	0,111	0,059	0,051	0,064	1,000	0,998	0,924	0,445	0,197	0,109	0,092	0,101

5 Conclusions and Future Works

In this paper we investigated the robustness of visual saliency methods against image corruption (additive noise and image compression). We compared the results of three popular saliency estimation methods with a ground truth that consists of real fixation maps. Several experiments have been conducted in order to analyze the effect of image corruptions in the resulting saliency maps, we focus our attention on additive noise (salt & pepper, gaussian noise) and jpeg compression.

As we expected, statistical accuracy measures of saliency maps decreases with respect to increasing global spatial distribution of noise into the images, GBVS method outperforms the other approaches showing accuracy values very close to the original values (i.e. the comparison between the real fixation maps and the saliency maps of the original images). Otherwise, the effects of jpeg compression depend on some features, such as the morphological surface of the image and the level of complexity of the scene. For this reasons we investigated on the performances of saliency detection methods and the rate of image compression: by observing preliminary results we noticed that jpeg compression substantially do not affect the performance of saliency detection methods.

It also would be interesting to analyze the relationship between the color quantization and the single steps of image compressions and the visual perception. In future works, we also want to extend this analysis to a larger set of images under varying conditions.

References

1. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapidscene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(11), 1254–1259 (1998)
2. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Advances in NeuralInformation Processing Systems, vol. 19, pp. 545–552. MIT Press (2007)
3. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiology **4**, 219–227 (1985)
4. Luo, J.: Subject content-based intelligent cropping of digital photos. In: IEEE International Conference on Multimedia and Expo (2007)
5. Sundstedt, V., Chalmers, A., Cater, K., Debattista, K.: Topdown visual attention for efficient rendering of task related scenes. In: Vision, Modeling and Visualization, pp. 209–216 (2004)
6. Itti, L., Koch, C.: Computational modeling of visual attention. Nature Reviews Neuroscience **2**(3) (2001)
7. Chen, L.-Q., Xie, X., Fan, X., Ma, W.-Y., Zhang, H.-J., Zhou, H.-Q.: A visual attention model for adapting images on small displays. ACM Multimedia Systems Journal **9**(4) (2003)
8. Judd, Y., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: IEEE 12th International Conference on Computer Vision, pp. 2106–2133 (2009)
9. http://people.csail.mit.edu/tjudd/WherePeopleLook/index.html
10. Gide, M.S., Karam, L.J.: Comparative evaluation of visual saliency models for quality assessment task. In: International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM) (2012)
11. Redi, J., Liu, H., Zunino, R., Heynderickx, I.: Interactionsof visual attention and quality perception. Proceedings of SPIE, Human Vision and ElectronicImaging XVI **7865**(1), 1–11 (2011)
12. Toet, A.: Computational versus psychophysical bottom-up image saliency: A comparative evaluation study. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(11), 2131–2146 (2011)
13. Duncan, K., Sarkar, S.: Saliency in images and video: a brief survey. IET Computer Vision **6**(6), 514–523 (2012)
14. Simoncelli, E.P., Freeman, W.T.: The steerable pyramid: A flexible architecture for multiscale derivative computation, pp. 444–447 (1995)
15. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision **42**, 145–175 (2001)
16. Rosenholtz, R.: A simple saliency model predicts a numberof motion popout phenomena. Vision Research **39**(19), 3157–3163 (1999)
17. Treisman, A.M., Gelade, G.: A Feature-Integration Theory of Attention. Cognitive Psychology **12**(1), 97–136 (1980)