An Experimental Analysis of Saliency Detection with respect to Three Saliency Levels

Antonino Furnari, Giovanni Maria Farinella, Sebastiano Battiato
{*furnari,gfarinella,battiato*} *@dmi.unict.it*

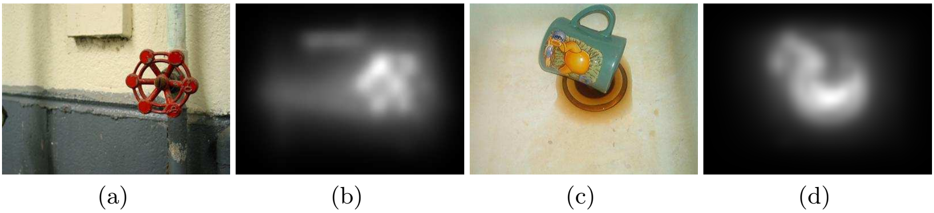
Department of Mathematics and Computer Science - University of Catania

Abstract. Saliency detection is a useful tool for video-based, real-time Computer Vision applications. It allows to select which locations of the scene are the most relevant and has been used in a number of related assistive technologies such as life-logging, memory augmentation and object detection for the visually impaired, as well as to study autism and the Parkinson's disease. Many works focusing on different aspects of saliency have been proposed in the literature, defining saliency in different ways depending on the task. In this paper we perform an experimental analysis focusing on three levels where saliency is defined in different ways, namely visual attention modelling, salient object detection and salient object segmentation. We review the main evaluation datasets specifying the level of saliency which they best describe. Through the experiments we show that the performances of the saliency algorithms depend on the level with respect to which they are evaluated and on the nature of the stimuli used for the benchmark. Moreover, we show that the eye fixation maps can be effectively used to perform salient object detection and segmentation, which suggests that pre-attentive bottom-up information can be still exploited to improve high level tasks such as salient object detection. Finally, we show that benchmarking a saliency detection algorithm with respect to a single dataset/saliency level, can lead to erroneous results and conclude that many datasets/saliency levels should be considered in the evaluations.

Keywords: saliency detection, visual attention modelling, salient object detection, salient object segmention, saliency levels, datasets for saliency evaluation

1 Introduction

During the last decades, we have observed the wide spread of affordable electronic devices capable of acquiring and processing images. This has virtually enabled a series of real-time Computer Vision applications which can rely on the large amount of data constantly gathered from the environment. Among these technologies, in particular, wearable devices provided with both computational power and a number of sensors (often including one or more cameras) are recently gaining more and more popularity. Since they involve egocentric vision,



(a) (b) (c) (d)

Fig. 1. Some sample images (a,c) and the related saliency maps (b,d).

wearable devices are rapidly changing the way we used to intend Computer Vision and are paving the way to a number of applications tightly coupled with the user's everyday life experience. Some of these applications are related to assistive technologies such as egocentric video summarization for life-logging [1] and memory augmentation [2, 3], object recognition for the visually-impaired [4], quality of life assessment and sensory substitution. Visual saliency has also been used for studying autism [5, 6] and Parkinson's disease [7].

In order to be able to manage all this incoming information in real-time, a mechanism able to select the parts of the image which are the most relevant with respect to the selected task, is needed to speed up the computation. Several studies argue that such a mechanism is likely to be present in the human system of attention [8–11]. Specifically, the human attentional phenomenon is believed to happen in at least two stages: 1) pre-attentive stage: is performed over the entire field-of-view (25 to 50 ms per item [9]) in order to select the locations which are sufficiently distinctive; 2) attentive stage: high level entities like objects are recognized through the combination of different features [12]. The first stage is an involuntary bottom-up process where the features automatically pop out according to their relationship with the surrounding (e.g., a red can on the grass is highly distinctive due to its red colour) [9, 13]. The second stage is a volitional top down process in which many different factors, including the subject's expectations (often related to the subject's knowledge of the scene) and the given task (e.g., free-viewing vs. object-search), are involved [9, 13]. These two stages of visual attention are usually modelled separately and tackled as different tasks, which gives rise to the distinction between bottom-up and top-down approaches. Methods aiming at exploiting both mechanisms fall into the class of integrated methods [14].

Building on this connection, Koch and Ullman [15] introduced the first biologically plausible model of attention, together with the concept of saliency map. A saliency map is a two dimensional topological map encoding the spatial locations conspicuousness, which can be directly exploited to select the most relevant regions of the scene (see Fig. 1 for some examples). An important result of such a connection is that weighting the importance of the acquired information using a biologically plausible model of attention allows a representation of the scene which is likely to be close to the human one. Many different approaches to visual saliency have been proposed through the last decades (see [13, 14] for comprehensive reviews) and some of them have been used to model visual attention in complex systems [4]. Some saliency models are biologically plausible [16–18], some are purely computational [19, 20], while others are mainly computational but still based on some biological cues [21–23].

Many different categorizations of the saliency detection methods are available in literature [13, 14], but most of the authors agree on the fact that visual attention depends on the task [12–14]. Evidences of this task-dependency date back to the seminal works on eye movements and vision performed by Yarbus in the late 60s [24]. Considering that saliency detection is an useful instrument which can be integrated into a wide variety of real-time, video-based applications [1–7], we argue that attention should be paid to the level with respect to which saliency is considered. Specifically we distinguish three different levels of saliency, namely Visual Attention Modelling (VAM), Salient Object Detection (SOD) and Salient Object Segmentation (SOS). In this paper we show that algorithms designed to deal with a specific level have different performances on the other levels. This fact has to be taken into account during the testing phase of a saliency detection method. We also show that the datasets used for the evaluations should be chosen carefully in order to properly assess the algorithms' performances with respect to one or more of the selected levels. For the evaluations, we select 8 relevant saliency detection algorithms [16, 19–23, 25] which we divide into the three aforementioned categories: visual attention modelling, salient object detection and salient object segmentation. We assess the performances of each method against different public datasets provided with different kinds of ground truth (eye fixation maps for visual attention modelling, bounding boxes for salient object detection and pixel-wise object masks for salient object segmentation). We show the results in the form of ROC curves [13] and AUC values [13] and compare them yielding a discussion.

The contributions of our work are the following: we review the most relevant datasets which can be used to evaluate the performances of saliency detection algorithms with respect to the considered levels; we show through experimental evidences that the performances of saliency detection algorithms depend on the considered saliency level and on the nature of the stimuli; we show that eye fixation maps can be effectively used to perform salient object detection and segmentation, which suggests that bottom-up cues are important even for higher level tasks as object detection; finally we show that a given algorithm should be evaluated with respect to different datasets/saliency levels in order to obtain correct evaluations.

The remainder of the paper is organized as follows: in Section 2 we discuss the three saliency levels we focus on in this paper; in Section 3 we present some related works and review the most relevant datasets with respect to the considered saliency levels; Section 4 defines the experimental settings and the used evaluation scores; whereas in Section 5 the results are discussed. Finally Section 6 concludes the paper and gives insights about further research.

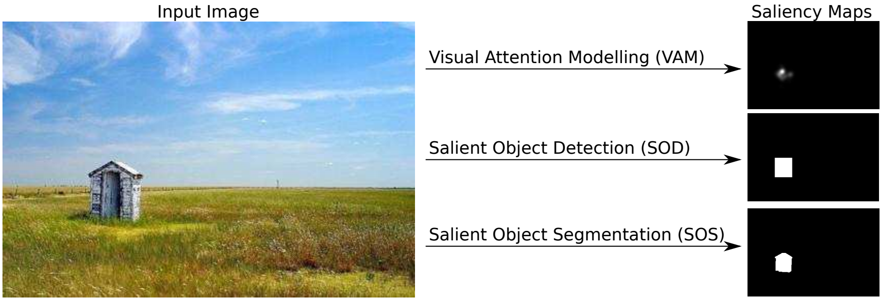

Input Image
Saliency Maps
Visual Attention Modelling (VAM)
Salient Object Detection (SOD)
Salient Object Segmentation (SOS)

Fig. 2. An example of three different levels at which saliency should be considered in designing a saliency detection algorithm. The input image is shown on the left, whereas the "ideal" saliency maps related to the three different levels are shown on the right.

2 Saliency Levels

As already mentioned we focus on three saliency levels, which are closely related to three different tasks. In the following we describe each level discussing which stages of the attentional phenomenon are mainly involved. Fig. 2 shows some examples of saliency maps which should be computed by an algorithm with respect to different levels. Since we argue that the datasets used for the evaluation should be differentiated according to the task, we also mention which type of ground truth is best used to evaluate the algorithms' performances with respect to the considered levels.

- **Visual Attention Modelling (VAM)** refers to the ability of the saliency map to predict the human eye fixations. This level is related to the pre--attentive stage where the most distinctive spatial locations are selected in a bottom-up manner, basing on the relationship of their features with the surroundings. To benchmark this ability, the saliency maps are compared with eye fixation maps (see Fig. 3 (b) for some examples) which are acquired by tracking the eye movements of many subjects looking at the stimuli [13]. No special task, but free viewing is generally assigned to the subjects;
- **Salient Object Detection (SOD)** refers to the ability of the saliency map to detect the salient objects in the scene. Some cues from both the pre--attentive (e.g., local features distinctiveness) and the attentive (e.g., prior knowledge about the object features) stages are generally involved in this task. Algorithms aiming at salient object detection are best benchmarked against datasets provided with bounding boxes (see Fig. 3 (c) for some examples) annotating for each stimulus the position of the objects [26];
- **Salient Object Segmentation (SOS)** refers to ability of the saliency map to detect and segment the salient objects present in the scene. Also in this case, the integration of the bottom-up and top-down levels is generally involved. The performances of salient object segmentation algorithms are generally assessed using datasets containing pixel-wise object masks [21] (see Fig. 3 (c) for some examples).

	Acronym	Study	Year	Level
1	IT	Itti et al. [16]	1998	VAM
2	IS	Hou et al. [20]	2011	VAM
3	GB	Harel et al. [22]	2007	VAM
4	AWS	Garcia et al. [23]	2012	VAM
5	SR	Hou et al. [19]	2007	SOD
6	CA	Goferman et al. [25]	2012	SOD
7	FT	Achanta et al. [21]	2009	SOS
8	CB	Jiang et al. [27]	2011	SOS

Table 1. The saliency detection algorithms used in the experiments. VAM = Visual Attention Modelling, SOD = Salient Object Detection, SOS = Salient Object Segmentation.

3 Related Work

Many saliency detection algorithms are available in the literature. Here we consider some of the available methods and organize them into the three considered levels (see Section 2) taking into account what stated by the authors in the related publications or the type of the ground truth data used to assess the algorithm's performances. Moreover we review the most relevant datasets introduced in the literature, associating each dataset with one or more of the discussed levels.

3.1 Saliency Methods

We consider 8 algorithms for our analysis. We chose them according to their popularity ([16, 19, 21]), the variety of their approaches ([16] is biologically plausible, [19, 20] are purely computational, [22] uses a probabilistic framework, [27] integrates object-level shape priors), and the performances exhibited in other benchmark papers ([22, 23, 27, 25] have good performances in [28, 29]).

The first computational model capable of producing saliency maps from input images was introduced in 1998 by Itti et al. [16]. Due to its biological plausibility, the model has been widely used as a benchmark for comparisons. In [22] Harel et al. introduce a saliency method which employs a graph-based probability model. Hou et al. have worked on spectral based approaches [19, 20] exploring the connections between information redundancy and the spectral content of the input image. In [23], Garcia et al. present a visual saliency method which relies on a contextually adapted representation produced through adaptive whitening of colour and scale features. In [25], Goferman et al. introduce an approach which aims at detecting the image regions that represent the scene, building on four principles observed in the psychological literature. In [21], Achanta et al. present an approach based on the analysis of the frequency content of the image. Jiang et al. [27] concentrate on salient object segmentation considering both bottom-up cues and object-based shape priors. Table 1 summarizes the algorithms considered in this paper with the related saliency level for which they have been designed.

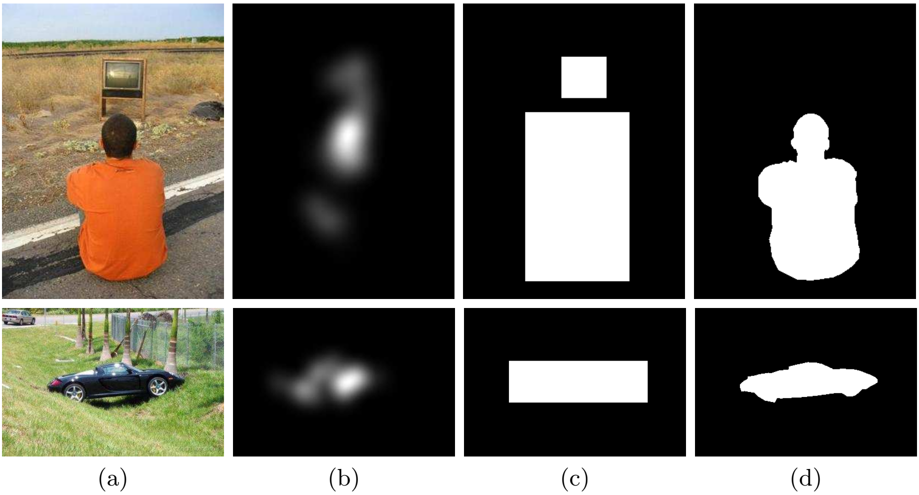


Fig. 3. Some examples of different types of Ground Truth. The stimuli (a), Eye Fixation Maps (b), Bounding Boxes annotations (c) and Pixel-Wise masks (d).

3.2 Datasets

In this subsection we review some of the most relevant datasets which have are used in the experiments. We analyse them with respect to different factors including the nature of the provided stimuli (e.g., natural images or images always containing a salient object), the nature of the ground truth and the way the ground truth is generated (e.g., how many labellers or subjects are employed). In particular we consider three types of ground truth:

- eye fixation maps, which are obtained using eye tracking data gathered from different subjects watching the stimuli;
- bounding boxes annotating the position of the salient objects of interest;
- pixel-wise masks of the salient objects depicted.

In Fig. 3 some examples of the different types of ground truth are provided. The considered datasets have been selected according to their popularity, the number of the stimuli, their diversity and the quality of the provided ground truth.

- **Microsoft Research ASIA Dataset (MSRA)** [26] is composed of 25000 images each containing a salient object of interest. The images are gathered from forums and image search engines. The dataset has been introduced for salient object detection and the ground truth consists in bounding boxes annotations. Each image is labelled by three users which are asked to draw a rectangle to specify a salient object. Considering the bounding boxes annotations, the dataset is related to the salient object detection level (SOD);
- **MIT Eye Fixations Dataset** [30] contains 1003 natural indoor and outdoor images which are viewed by 15 subjects in order to obtain eye fixations data. No particular instruction but free-viewing was given to the observers.

	Dataset	**Study**	**Year**	**Images**	**Ground Truth**	**Level**
1	MSRA	Liu et al. [26]	2011	10000	BB	SOD
2	MIT	Judd et al. [30]	2009	1003	EF	VAM
3	THUS10000	Cheng et al. [31]	2011	10000	PW	SOS
4	DUT-ORMON	Yang et al. [32]	2013	5168	All	All

Table 2. The considered datasets. EF = Eye Fixation Maps (corresponding to the visual attention modelling level), BB = Bounding Boxes (corresponding to the salient object detection level), PW = Pixel Wise Masks (corresponding to the salient object segmentation level).

The saliency level related to this dataset is the visual attention modelling (VAM);

- **THUS10000 Datasets [31]** is derived from the MSRA [26] dataset by picking a subset of 10000 images. Each image is hand labelled at pixel accuracy level by a single subject in order to obtain a mask of the salient object depicted. The dataset is suitable for assessing the performances of the algorithms with respect to the level of salient object segmentation (SOS);
- **DUT-ORMON Dataset [32]** contains 5168 high quality images. Three kinds of ground truth are provided with this dataset: 1) the pixel-wise masks of the salient objects, 2) the bounding boxes annotations of the salient objects and 3) the eye fixations data. The instructions given to the labellers are similar to the ones given for the other datasets. The ground truth is built employing 5 participants per image. Considering the three types of ground truth provided, the dataset is suitable for all the three considered saliency levels: visual attention modelling (VAM), salient object detection (SOD) and salient object segmentation (SOS);

Table 2 reports a summary of the considered datasets, providing information on the number of images, the type of the ground truth and the related levels.

3.3 Other Saliency Benchmark Works

At least two other saliency benchmark papers are related to the present work. In [28] the authors examine several salient object detection approaches with respect to different salient object detection and segmentation datasets. They also discuss the dataset properties, the evaluation measures to be used and the effects of the aggregation of different saliency methods. In [29] the same authors benchmark several fixation prediction algorithms against different eye fixations datasets, discussing which measures are best used for such evaluations.

Differently than the works in [28, 29], which analyse the two main levels of saliency (salient object detection and visual attention modelling) separately, in this paper we compare the results of the selected algorithms with respect to different types of ground truth. In addition, we take advantage of the DUT-ORMON dataset introduced in [32] which contains different types of ground truth for the same stimuli in order to yield more consistent and comparable evaluations of the saliency algorithms considered in this paper.

4 Experimental Settings

For the experiments we considered the 8 algorithms which have been presented in Section 3 and summarized in Table 1 and the 4 datasets which have been reviewed in Section 3.2 and reported in Table 2. In order to produce saliency maps for the evaluations, we used the original code provided by the authors, which is publicly available. For sake of fairness we don't tune the parameters of the algorithms, hence using the standard ones which are provided by the authors.

The first dataset we consider is the DUT-ORMON [32] dataset. It is useful to compare the performances of algorithms with respect to the three different saliency levels, since it contains all the three types of ground truth for the same stimuli. The second one is the MIT [30] eye fixations dataset, which we use to benchmark the algorithms with respect to the level of visual attention modelling. The third one is the MSRA [26] dataset, which has been introduced to evaluate salient object detection algorithms and contains bounding boxes annotations for each image. The fourth dataset is the THUS10000 [31] dataset which contains 10000 images taken from the MSRA dataset but provides pixel-level ground truth and is suitable to evaluate salient object segmentation algorithms. Since the THUS10000 dataset is derived from MSRA by picking a subset of the provided stimuli, for comparison sake, when benchmarking against the MSRA dataset, we consider that subset of stimuli and refer to this modified dataset as MSRA10000.

We perform three comparative experiments aimed at analysing the performances of the saliency detection algorithms at the different levels. In particular we perform tests to assess: 1) the performances of the algorithms with respect to the same stimuli but different levels, 2) the performances of the algorithms with respect to the same level, but different stimuli, and 3) to what degree there's a connection among the different levels. The final aim of this work is to show that the different levels of saliency should be taken into account when evaluating the algorithms. Consequently we want to show that considering a single saliency level or a single dataset (which is a common practice in the literature) can lead to erroneous evaluations. For each experiment we provide ROC curves and AUC scores as described in [13, 14, 20].

Experiment 1: in the first experiment we assess the performances of the considered algorithms with respect to the DUT-ORMON [32] dataset and its different types of ground truth. The evaluations produced by this experiments are comparable with respect to the different levels, since the stimuli are the same (i.e., the saliency maps are computed only once and benchmarked against the different types of ground truth). Nevertheless we expect the relative ranking among the algorithms could change; some algorithms should perform better on some levels than on some others;

Experiment 2: in the second experiment we evaluate the performances of the algorithms with respect to the MIT eye fixations, THUS10000 and MSRA10000 datasets. We compare the results with respect to the performances of the algorithms on the corresponding levels of the DUT-ORMON dataset. Even if the considered saliency levels are the same, we expect some different results with respect to Experiment 1, since the nature of the stimuli is different in some cases;

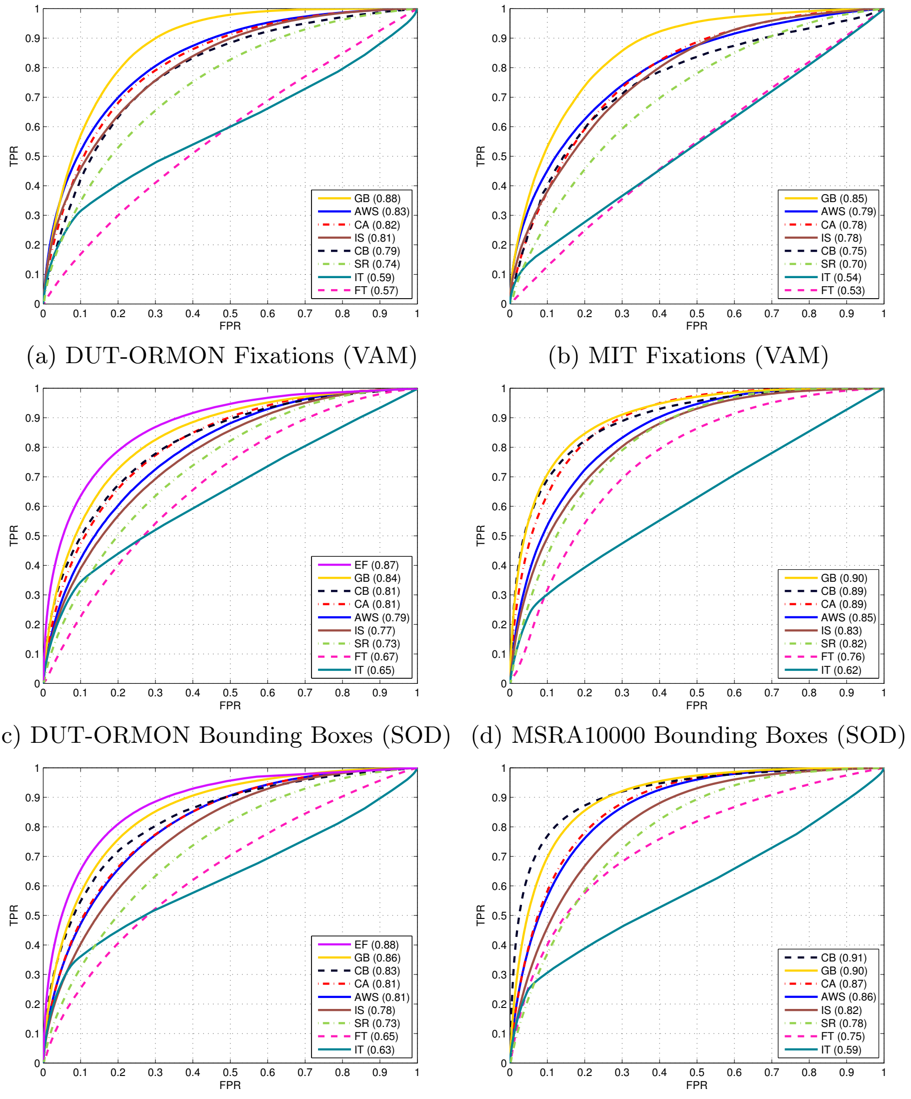
Experiment 3: in the third experiment, we consider the eye fixations maps included in the DUT-ORMON dataset as saliency maps and evaluate their performances on both salient object detection and segmentation on the same dataset. Since the eye fixation maps are likely to be sparse, in order to be fair with respect to the other saliency detection algorithms, we first convolve each eye fixation map with a Gaussian kernel with variance $\sigma = 15\ pixels$. We compare the results with the performances of the other algorithms on the same dataset. This experiment tells if there is a strong connection among the different levels (e.g., if an algorithm designed for eye fixations can detect salient objects) or if the two tasks should be considered independent (e.g., if an algorithm designed for eye fixations cannot be used to detect salient objects).

5 Results & Discussion

The ROC curves related to the three experiments proposed in Section 4 are reported in Fig. 4. For each diagram, the AUC values related to the ROC curves are reported in parenthesis in the legend which is sorted in descending order to assess the algorithms' ranking. The ROC curves line styles are related to the algorithms' saliency levels: solid line for VAM, dash-dot line for SOD, dashed line for SOS. The first column of Fig. 4 contains the ROC curves related to the performances of the algorithms on the DUT-ORMON dataset with respect to different levels of saliency: visual attention modelling (a), salient object detection (c) and salient object segmentation (e). In the diagrams (c) and (d), the performances of the considered algorithms are also compared to the performances of the eye fixation maps when they are used to perform object detection and segmentation. These additional ROC curves are referred to in the legend as the "EF" series. The second column of Fig. 4 reports the ROC curves for the experiments performed on the other datasets.

In Fig. 5 a comparative diagram of the results is reported. The diagram visualizes the AUC values for each algorithm in all the evaluation settings (referred to in the form "Dataset - Level"). The line styles refer to the algorithms' saliency levels as in Fig. 4. The overall performances of the considered algorithms on all the experimental settings are measured computing the normalized area under the comparative curves through numerical integration. Those values are reported in parentheses in the legend which is sorted in descending order.

Fig. 6 shows some sample saliency maps computed on images taken from the fourth considered datasets for visual assessment. For each image the saliency maps computed by all the considered algorithms and the related ground truth data are reported.



(a) DUT-ORMON Fixations (VAM)

(b) MIT Fixations (VAM)

(c) DUT-ORMON Bounding Boxes (SOD)

(d) MSRA10000 Bounding Boxes (SOD)

(e) DUT-ORMON Pixel-Wise (SOS)

(f) THUS10000 Pixel-Wise (SOS)

Fig. 4. The ROC curves for the performed experiments. The ROC line styles refer to the algorithms' saliency levels: solid line for VAM, dash-dot line for SOD, dashed line for SOS. The corresponding AUC scores are reported in parenthesis in the legends. Each legend is sorted by AUC score in descending order to highlight the ranking of the algorithms in the considered settings. The first column (a, c, e) reports the performances of the same algorithms with respect to different saliency levels on the DUT-ORMON dataset. Each row (a-b, c-d, e-f) compares the performances of the same algorithms with respect to the same level but different datasets. The effects of performing salient object detection and segmentation using the ground truth fixation maps on the DUT-ORMON dataset, are reported in (c) and (e) and refer to series "EF".

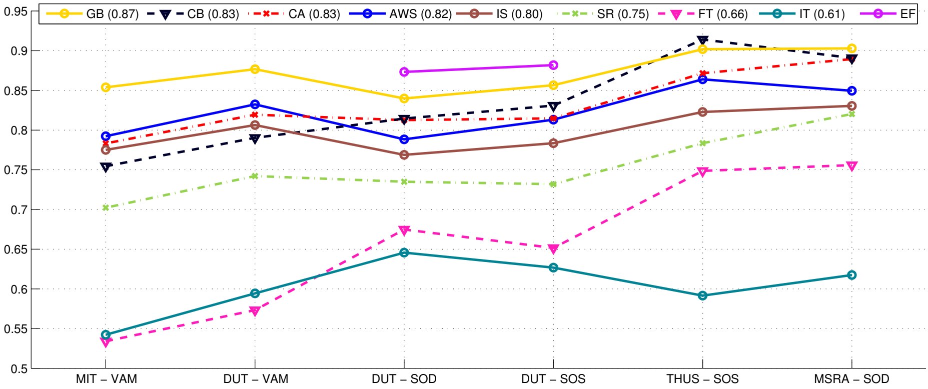


Fig. 5. The diagram visualizes the AUC evaluations of the different algorithms whit respect to different dataset/ground truth combinations. The line styles refer to the algorithms' saliency levels: solid line for VAM, dash-dot line for SOD, dashed line for SOS. The "EF" series represents the performances of the eye fixations maps on the SOD and SOS level in the DUT-ORMON dataset. As an overall evaluation among the different evaluation settings, the normalized areas under the shown curves are reported in parenthesis in the legend, which is sorted in descending order. No overall evaluation of the "EF" series is provided since it is not extended to all the experimental configurations.

5.1 Discussion on Experiment 1

In Fig. 5, looking at the transitions between the "DUT-VAM" configuration, the "DUT-SOD" configuration and the "DUT-SOS" configuration, it can be noted that the ranking of the algorithms generally changes according to the saliency level. Since in the considered experimental configurations each algorithm is benchmarked against the same stimuli, we can assert that the change of ranking is entirely due to the different levels at which saliency is defined by the different types of ground truth. The ranking related to salient object detection ("DUT-SOD") and salient object segmentation ("DUT-SOS") is unchanged and the AUC scores slightly change according to the algorithms. This suggests that the SOD and SOS levels are closely related. A further analysis could be aimed at assessing this intuition in a more rigorous way. Moreover, it can be noted that some algorithms which are tailored to object detection or segmentation (e.g., CB, FT, SR and CA) have higher (or similar) performances in the SOS and SOD levels than in the VAM level. Whereas, the algorithms tailored to the VAM level (e.g., GB, AWS, IS) have generally lower performances on the SOS and SOD levels. The IT algorithm is the only exception to this scheme, giving better results on the SOD and SOS levels even if tailored to the VAM level. The GB algorithm has the best performances with respect to all the considered saliency levels on the DUT-ORMON dataset, which means that it is capable at the same time of predicting the eye fixations and performing salient object detection and segmentation. Some examples are available in Fig. 6 for visual assessment.

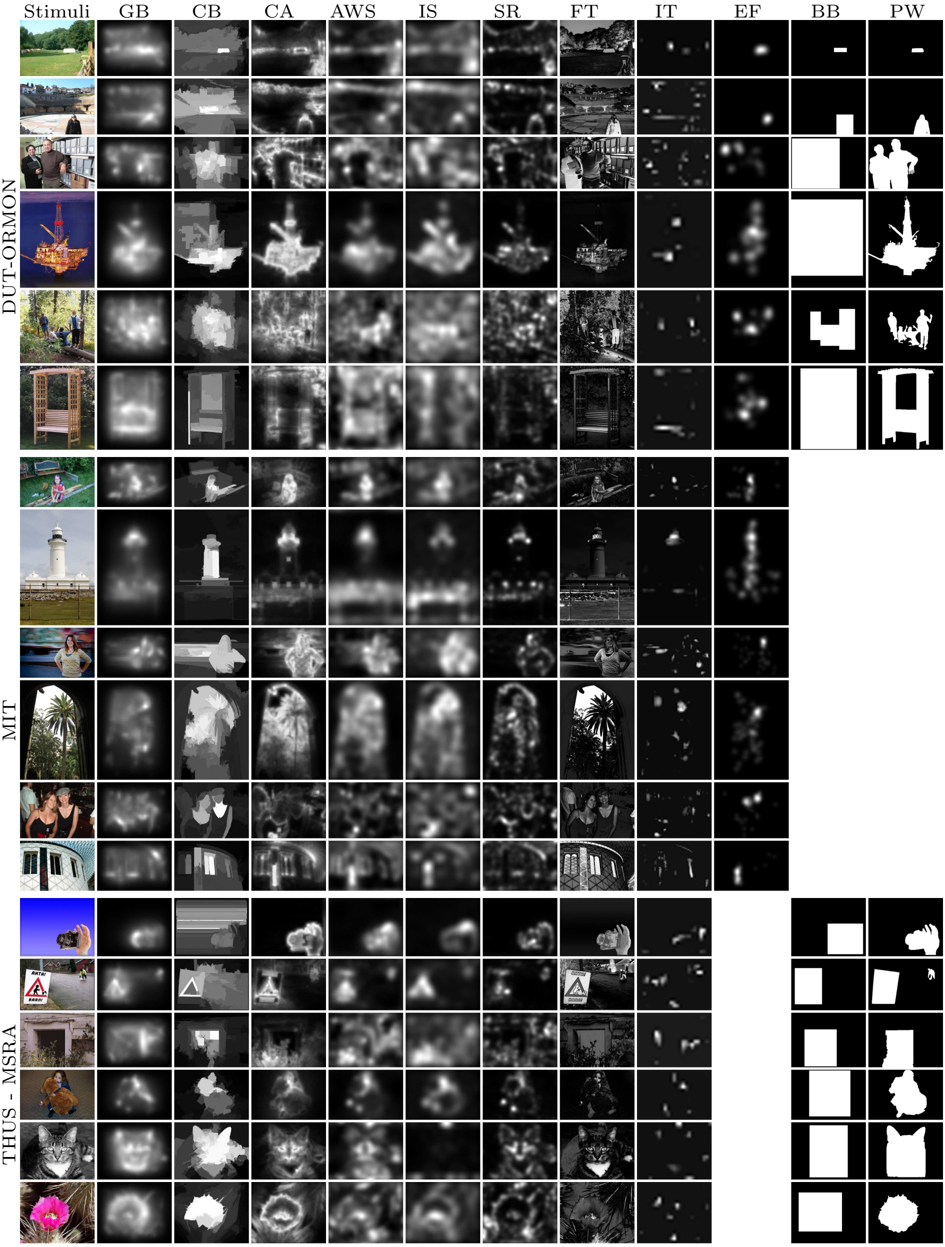


Fig. 6. Some sample saliency maps computed by the considered algorithms. Six images (first column) with the corresponding ground truths (last three columns) are reported for each dataset. The THUS and MSRA datasets are grouped since the stimuli are the same. EF = Eye Fixations maps, BB = Bounding Boxes annotations, PW = Pixel-Wise masks.

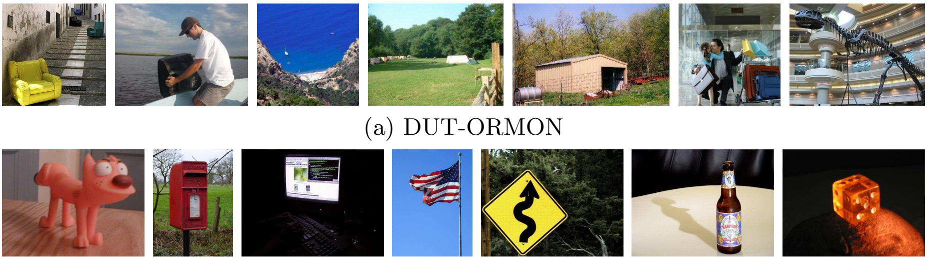
5.2 Discussion on Experiment 2

The transition between the “DUT-SOS” and the “THUS-SOS” configurations in Fig. 5, shows that most of the algorithms perform much better on the THUS1000 dataset. The only exception is the IT algorithm whose performances are lower on the THUS0000 dataset. The transition between the “DUT-SOD” and the “MSRA-SOD” datasets shows a general increase in the performances except for the IT aglorithm. Moreover the CB algorithm performs better than GB on the THUS10000 dataset, even if it performed worse with respect to the same level on the DUT-ORMON dataset (see Experiment 1). In our opinion, this change of ranking is due to the different nature of the stimuli contained in the two datasets. The THUS dataset has been derived from the MSRA dataset, which was introduced for the task of salient object segmentation, so the stimuli often contain a clearly distinguishable salient object of interest. The stimuli of the DUT-ORMON dataset are more general (and hence more “difficult”) and don’t always contain a clearly distinguishable salient object of interest. A visual example of this statement is provided in Fig. 7. Since the CB algorithm has been explicitly designed for salient object segmentation and makes use of context and shape prior, it is likely to work better when this information can be exploited (see Fig. 6 for visual assessment). This leads to the conclusion that, if the tackled task is the salient object detection in very controlled conditions (where the object is clearly distinguishable from its context), the THUS dataset is probably good for the evaluations, while the DUT-ORMON dataset is a more challenging dataset which can be used to assess the performances of an algorithm which is designed to work in less constrained settings.

The transition between the “MIT-VAM” and “DUT-VAM” shows that the AUC values for the eye fixations in the DUT-ORMON are consistent with (and generally higher than) the values obtained on the MIT dataset. This is justified by the fact that for both datasets the task assigned to the observer was free-viewing and the stimuli included in the two datasets are similar (natural images with variable contexts). Moreover, this finding emphasizes that predicting the eye fixations is a “less ambiguous” task than predicting the salient objects of interest, where other factors like the subject knowledge or the context of the scene are involved. Moreover the increase of the results in the “DUT-VAM” configurations underlines that the MIT dataset is generally more challenging for the visual attention modelling level.

5.3 Discussion on Experiment 3

The series “EF” in Fig. 5 shows the performances of the eye fixation maps contained in the DUT-ORMON dataset when used to perform salient object detection and salient object segmentation on the same dataset. It can be observed that the eye fixations are suitable for detecting and segmenting the salient objects of interest, performing better than all the considered algorithms. The performances of the eye fixation can be considered as an upperbound to the performances of algorithms designed for predicting the eye fixations when applied to the other levels of saliency. This leads us to the conclusion that there is more room for improving the results of the salient object detection/segmentation algorithms



(a) DUT-ORMON

(b) THUS10000

Fig. 7. Some images from the DUT-ORMON dataset (a) and from the THUS10000 (b) dataset. As can be noted, the THUS10000 images always contain a clearly distinguishable salient object of interest, while most of the images from the DUT-ORMON dataset contain natural images where the salient objects are not always easily distinguishable.

still relying on low level bottom-up cues. However it should be considered that the DUT-ORMON dataset contains natural images, where the salient objects are not always clearly distinguishable or present, and so they are more likely to correspond with the eye fixations. It would be certainly interesting to assess the performances of the eye fixations on a dataset conceived for the object detection task like the MSRA dataset, but, considering that this data is not available, this is out of the scope of the present work.

6 Conclusion

We have studied saliency at three different levels, namely visual attention modelling, salient object detection and salient object segmentation. We have shown through experimental evidence that the performances of the algorithms generally depend on the saliency level with respect to which they are benchmarked. Comparing the performances of the algorithms with respect to datasets provided with different stimuli but same type of ground truth, we have assessed that the performances also depend on the nature of the stimuli. In particular, a closer look to Fig. 5 would reveal that using a single dataset can lead to erroneous evaluations (e.g., THUS-SOS vs DUT-SOS/DUT-SOD). We have noted that visual attention modelling is a "less ambiguous task", since the evaluations agree for different eye fixation datasets. We have shown that there is a strong relationship between visual attention modelling and salient object detection and segmentation, since the eye fixation maps can be successfully used for such tasks. Considering that the eye fixation maps yield the best results with respect to the considered algorithms, we argue that there is more room for improving object detection and segmentation still relying on bottom-up pre-attentive information. Moreover, we have reviewed the main datasets available in the literature, associating them to the analysed saliency levels according to the provided ground truths.

Future works will be devoted to extend our analysis to a larger number of saliency detection algorithms and datasets. We will study how saliency algorithms perform with respect to datasets composed of search arrays and psychological patterns [19]. Moreover, the level of visual attention modelling will be studied in both the dynamic and the static domain as suggested in [33].

References

1. Ghosh, J., Grauman, K.: Discovering important people and objects for egocentric video summarization. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012) 1346–1353
2. Hodges, S., Berry, E., Wood, K.: SenseCam: a wearable camera that stimulates and rehabilitates autobiographical memory. Memory **19**(7) (2011) 685–96
3. Lu, Z., Grauman, K.: Story-driven summarization for egocentric video. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2013) 2714–2721
4. Thakoor, K.A., Marat, S., Nasiatka, P.J., McIntosh, B.P., Sahin, F.E., Tanguay, A.R., Weiland, J.D., Itti, L.: Attention biased speeded up robust features (AB-SURF): A neurally-inspired object recognition algorithm for a wearable aid for the visually-impaired. IEEE International Conference on Multimedia and Expo Workshops (ICMEW) (2013) 1–6
5. Freeth, M., Foulsham, T., Chapman, P.: The influence of visual saliency on fixation patterns in individuals with autism spectrum disorders. Neuropsychologia **49**(1) (2011) 156–160
6. Amso, D., Haas, S., Tenenbaum, E., Markant, J., Sheinkopf, S.J.: Bottom-up attention orienting in young children with autism. Journal of autism and developmental disorders **44**(3) (2014) 664–673
7. Mannan, S.K., Hodgson, T.L., Husain, M., Kennard, C.: Eye movements in visual search indicate impaired saliency processing in parkinson's disease. Progress in brain research **171** (2008) 559–562
8. Ungerleider, S.K., G., L.: Mechanisms of visual attention in the human cortex. Annual review of neuroscience **23**(1) (2000) 315–341
9. Itti, L., Koch, C.: Computational modelling of visual attention. Nature reviews. Neuroscience **2**(3) (2001) 194–203
10. Rolls, E.T., Deco, G.: Attention in natural scenes: neurophysiological and computational bases. Neural networks **19**(9) (2006) 1383–1394
11. Koch, K., McLean, J., Segev, R., Freed, M.A., Berry II, M.J., Balasubramanian, V., Sterling, P.: How much the eye tells the brain. Current Biology **16**(14) (2006) 1428–1434
12. Treisman, A., Gelade, G.: A feature-integration theory of attention. Cognitive psychology **136**(1) (1980) 97–136
13. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. IEEE transactions on pattern analysis and machine intelligence (PAMI) **35**(1) (2013) 185–207
14. Duncan, K., Sarkar, S.: Saliency in images and video: a brief survey. IET Computer Vision **6**(6) (2012) 514–523
15. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Human neurobiology **4**(4) (1985) 219–227
16. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **20**(11) (1998) 1254–1259
17. Le Meur, O., Le Callet, P., Barba, D., Thoreau, D.: A coherent computational approach to model bottom-up visual attention. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **28**(5) (2006) 802–817
18. Kootstra, G., Nederveen, A., De Boer, B.: Paying attention to symmetry. In: Proceedings of the British Machine Vision Conference (BMVC). (2008) 1115–1125

19. Hou, X., Zhang, L.: Saliency detection: A spectral residual approach. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (800) (2007) 1 – 8
20. Hou, X., Harel, J., Koch, C.: Image Signature: Highlighting Sparse Salient Regions. IEEE transactions on pattern analysis and machine intelligence (PAMI) **34**(1) (2011) 194–201
21. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2009) 1597–1604
22. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. Advances in neural information processing systems **19** (2007) 545
23. Garcia-Diaz, A., Fdez-Vidal, X.R., Pardo, X.M., Dosil, R.: Saliency from hierarchical adaptation through decorrelation and variance normalization. Image and Vision Computing **30**(1) (2012) 51–64
24. Yarbus, A.L., Haigh, B., Rigss, L.A.: Eye movements and vision. Volume 2. (1967)
25. Goferman, S., Zelnik-Manor, L., Tal, A.: Context-aware saliency detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **34**(10) (2012) 1915–1926
26. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., Shum, H.Y.: Learning to detect a salient object. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **33**(2) (2011) 353–367
27. Jiang, H., Wang, J., Yuan, Z., Liu, T., Zheng, N., Li, S.: Automatic salient object segmentation based on context and shape prior. Volume 3. (2011) 7
28. Borji, A., Sihite, D.N., Itti, L.: Salient object detection: A benchmark. In: European Conference on Computer Vision (ECCV). (2012) 414–429
29. Borji, A., Sihite, D.N., Itti, L.: Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. IEEE Transactions on Image Processing (TIP) **22**(1) (2013) 55–69
30. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: IEEE International Conference on Computer Vision (ICCV). (2009)
31. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H.S., Hu, S.M.: Salient object detection and segmentation. Technical report, Tsinghua University (2011)
32. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2013) 3166–3173
33. Nguyen, T.V., Xu, M., Gao, G., Kankanhalli, M., Tian, Q., Yan, S.: Static saliency vs. dynamic saliency: a comparative study. In: ACM international conference on Multimedia. (2013) 987–996