https://doi.org/10.1007/s42979-020-00204-0

ORIGINAL RESEARCH

Saliency from High-Level Semantic Image Features

Aymen Azaza[1,2] · Joost van de Weijer[2] · Ali Douik[1] · Javad Zolfaghari[2] · Marc Masana[2]

Received: 10 April 2020 / Accepted: 22 May 2020 / Published online: 11 June 2020

Abstract
Top-down semantic information is known to play an important role in assigning saliency. Recently, large strides have been made in improving state-of-the-art semantic image understanding in the fields of object detection and semantic segmentation. Therefore, since these methods have now reached a high-level of maturity, evaluation of the impact of high-level image understanding on saliency estimation is now feasible. We propose several saliency features which are computed from object detection and semantic segmentation results. We combine these features with a standard baseline method for saliency detection to evaluate their importance. Experiments demonstrate that the proposed features derived from object detection and semantic segmentation improve saliency estimation significantly. Moreover, they show that our method obtains state-of-the-art results on (FT, ImgSal, and SOD datasets) and obtains competitive results on four other datasets (ECSSD, PASCAL-S, MSRA-B, and HKU-IS).

Keywords Saliency · Object detection · Semantic segmentation

Introduction

Saliency is the quality of objects that makes them stand out with respect to others, thereby grabbing the attention of the viewer. Computational saliency can be roughly divided in three main research branches. Firstly, it is originally defined as a task of predicting eye-fixations on images [11]. Secondly, researchers use the term to refer to salient object estimation or salient region detection [6, 35, 65]. Here, the task is extended to identify the region, containing the salient object, which is a binary segmentation task for salient object extraction. Thirdly, more recently researchers on convolutional neural networks have also used the term of saliency map to refer to the activations of certain intermediate layers of the network. The focus in this paper is on salient object estimation, and we do not perform fixation map prediction, nor study the activation maps of neural networks.

Computational salient object detection aims to detect the most attractive objects in the image in a manner which is coherent with the perception of the human visual system. Visual saliency has a wide range of applications such as image retargeting [15], image compression [51], and image retrieval [61].

Initially, most saliency models were bottom-up approaches which are based on low-level features which are merged using linear and nonlinear filtering to get the final saliency map [6, 9]. Itti et al. [22] propose one of the first models for computational visual saliency which is based on the integration theory of Treisman [52] and uses several low-level bottom-up features including color, orientation, and intensity. Even though this method has been surpassed on popular baselines by many approaches, a recent study which optimized all its parameters found that it could still obtain results comparable to state-of-the-art [17]. Yang et al. [58] improve low-level features by considering their contrast with respect to the boundary of the image. Here, the boundary is used to model the background. Then, the saliency map is computed using graph-based manifold ranking. Perazzi et al.

✉ Aymen Azaza
aymen.azaza@cvc.uab.es

Joost van de Weijer
joost@cvc.uab.es

Ali Douik
alidouik@gmail.com

Javad Zolfaghari
jzolfaghari@cvc.uab.es

Marc Masana
mmasana@cvc.uab.es

[1] National Engineering School of Sousse, University of Sousse, Pole technologique de Sousse, Sousse, Tunisia

[2] Computer Vision Center, Barcelona, Spain

[47] apply a Gaussian filtering framework which is based on computing regional contrast and element color uniqueness to rank the saliency of regions.

Top-down approaches consider that high-level semantic understanding of the image plays an important role in saliency assignment. These methods first identify a subset of high-level concepts, such as faces, text, and objectness, which are detected in the image, and in a subsequent phase are used to compute the saliency map. The first set of papers on this subject concentrated on a limited set of semantic classes. Cerf et al. [9] add a face detector to their saliency approach. Ehinger et al. [12] compute saliency by combining scene context features, target features, and location. Judd et al. [23] consider the detection of faces, people, text, body parts, and animals to improve saliency estimation. Borji et al. [6] also include features based on the detection of face and text in their saliency estimation method. Recently, some methods have considered a wider range of classes for saliency detection, by also incorporating object detection or semantic segmentation results in the saliency pipeline [28, 44, 53, 59, 63]. All of these methods show that adding high-level semantic features to saliency computation improves results significantly.

Convolutional neural networks (CNN) [26, 32] have significantly improved the state-of-the-art of high-level image understanding. Instead of separately designing hand-crafted features and optimal classifiers for computer vision problems, these networks propose to learn end-to-end, optimizing both the feature representation and the classifier at the same time. These techniques have led to impressive performance gains in semantic image understanding. For example, the results for object detection on the popular PASCAL VOC 2010 dataset have improved from 29.6 [16] in 2010 to 68.8 (mAP) with fast R-CNN in 2015 [20]. Impressive improvements can also be seen for semantic segmentation on PASCAL VOC 2011 from 47.6 in 2012 [8] to 62.7 (mIoU) with fully convolutional networks in 2015 [33]. Given this large improvement in performance, we think it is timely to revisit top-down high-level features for saliency.

Given the significant improvements of high-level object detection and semantic segmentation, in this paper we aim to evaluate the impact of these high-level methods on the task of saliency estimation. Figure 1 shows an example of the importance of high-level features for saliency. As discussed above, high-level semantic information plays an important role when attributing saliency [9, 43]. In addition, a recent article titled “Where should saliency models look next?” [4] concluded that models continue to miss semantically meaningful elements in scenes. Our paper has the following contributions:

- We evaluate if knowledge of semantic classes in the image (from a variety of object groups including humans, vehicles, indoor and animals) can be used for better saliency estimation.
- We evaluate this based on two methods for high-level image understanding, namely object detection and semantic segmentation.
- We propose several new saliency features based on the high-level information coming from the object detection and semantic segmentation methods.
- We perform an extensive analysis on several standard datasets and evaluate the gain which is obtained by having access to this high-level information.

The organization of the paper is as follows: in Sect. 2, we present the related work. In Sect. 3, we give an overview of the proposed method. In Sect. 4, we describe the features computed from object detection, segmentation results and object proposals. Next, we provide details on

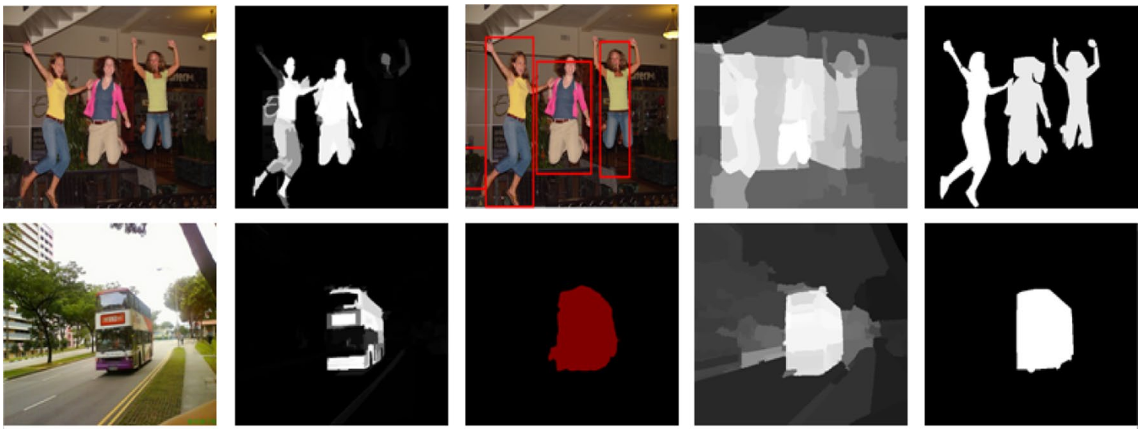


Fig. 1 From left to right (top row): input image, saliency map by MDF method [37], object detection results, our saliency map and ground truth; (bottom row): input image, saliency map by MDF method [37], semantic segmentation results, our saliency map and ground truth. Examples show that high-level features are important for saliency detection

the experimental setup and results are presented in Sect. 5. Conclusions are provided in Sect. 6.

Related Work

In this section, we provide an overview of salient object detection methods. After the seminal work of Itti et al. [22], who propose one of the first computational saliency models, saliency estimation has led to both biologically inspired models [5], and many mathematical motivated methods [2, 21]. A complete review on saliency can be found in [7].

Object proposal methods became a hot topic of research, due to their success in object detection [10, 25, 54, 64]. Recently, these object proposals methods have been applied in the field of saliency detection. The advantage of using object proposals approaches over methods based on superpixels is that they do not require an additional regrouping step (often implemented with a conditional random field [39]). The use of object proposal methods has another advantage which is avoiding the use of the costly sliding window approaches. Several methods use object proposals for saliency estimation [3, 27, 35, 62]. They extract saliency features for all object proposals after which they use a classifier to assign saliency to the object proposals. Recently, Azaza et al. [3] propose a saliency approach based on saliency features computed from the direct surround (context) of every object proposal. Wang et al. [62] propose a local and global deep network for saliency to predict the saliency of each object proposal generated from the geodesic object proposal method [25]. A recent work investigates the usage of object proposals for saliency estimation in videos [27].

Several methods have explicitly used high-level object detection for saliency estimation. Xu et al. [56] introduce a visual saliency approach which includes semantic attributes which are related to emotion, touch, gaze, smell, and taste. Nuthmann et al. [46] prove that objects are important in leading attention. Einhauser et al. [13] demonstrate that objects predict fixations better than early saliency, so they propose a model based on detecting or segmenting objects to predict salient regions. Other than these methods, we consider a wider group of twenty object classes and evaluate their impact on saliency estimation. In addition, we evaluate both the influence of object detection and semantic segmentation for saliency estimation.

Recently, the use of high-level semantic information (object detection and semantic segmentation) for saliency detection has been investigated in some papers [28, 44, 53, 63]. Hou et al. [28] combined multi-level feature maps of fully convolutional neural networks (FCNs), on which most object detection and semantic segmentation algorithms are based, to highlight the salient objects. Ming et al. [44] present a saliency approach using active semantic segmentation, which they use to extract a set of semantic features. Tingtian et al. [53] propose a saliency algorithm which merges high-level foreground object detection with low-level features. Wang et al. [63] propose a saliency detection approach based on a background prior method by merging three saliency maps obtained by background contrast, background connectivity prior, and spatial distribution prior. We differ from these works [28, 44, 53, 59, 63] in that we provide a more complete analysis of the influence of top-down information for saliency estimation: we analyze the relative gain of object detection and semantic segmentation results, and we analyze the relative gain provided by each of the semantic classes.

Recently, CNNs have been applied to visual saliency research. Initially, several works used off-the-shelf deep features to replace previous hand-crafted features [37, 49, 62]. Further progress was made when fully convolutional networks allowed for end-to-end estimation of saliency [38], which led to convolutional features optimized for saliency detection. Li et al. [30] propose a multi-task deep model for semantic segmentation and saliency prediction. They investigate the correlation between the semantic segmentation and saliency detection. They prove that using features collaboratively for two correlated tasks can improve overall performance. In this work, we study the influence of state-of-the-art semantic image understanding methods, such as object detection and semantic segmentation, on saliency detection. We use a standard baseline which is not based on deep learning, but the method could potentially be extended to include bottom-up deep features.

Method Overview

The main aim of this paper is to analyze the usage of high-level semantic information (object detection, and semantic segmentation results) for saliency estimation. To evaluate the impact of high-level semantic information on saliency, we use a standard saliency pipeline. A similar approach was for example used by Li et al. [30] where they propose a multi-task deep model for semantic segmentation and saliency prediction task.

An overview of the baseline saliency approach at testing time is shown in Fig. 2. Given an image, we compute a set of object proposals using the multiscale combinatorial grouping (MCG) method [1]. Based on the extracted feature vector for each of the object proposals, we train a random forest for regression to produce a saliency model which will be used for saliency estimation. As the saliency score for each object proposal, we use the average saliency of the pixels in the proposal (pixels have a saliency of one if they are on the ground truth salient object or zero elsewhere). At testing time, we assign saliency for all the object proposals

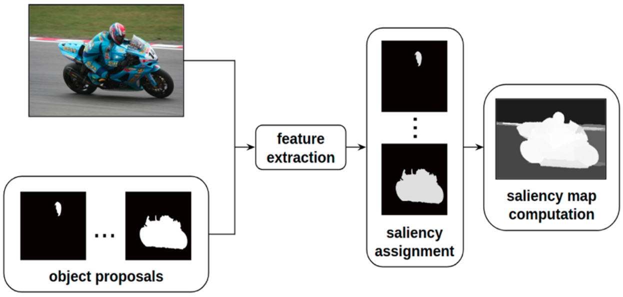


Fig. 2 Overview of our proposed method: from the input image, we compute a set of object proposals. From these objects, we compute shape, object detection, and segmentation features. Next, we train a random forest to classify the features of each proposal to a saliency assignment value. The saliency values of all proposals are combined in a saliency map

using the random forest regressor. The final saliency map is computed by taking for each pixel the average of the saliency of all the proposals that contain that pixel.

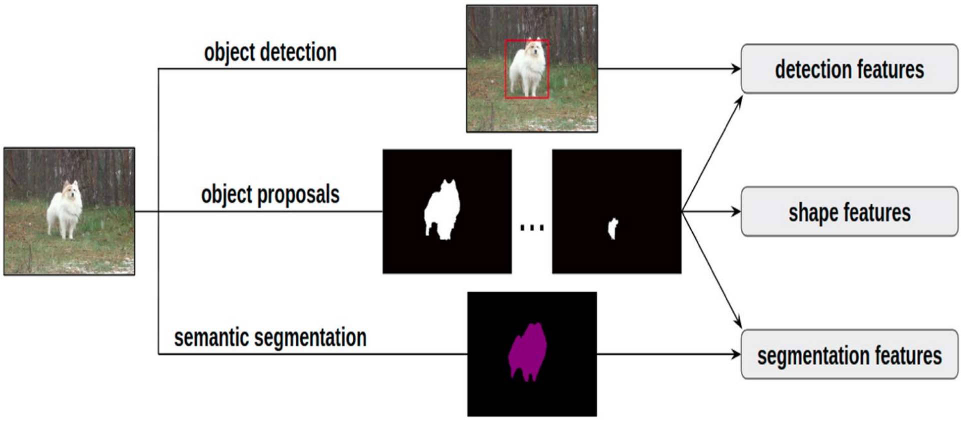
To incorporate high-level semantic information into the saliency pipeline, we only change the feature extraction phase of the baseline method. An overview is provided in Fig. 3. We will consider two types of high-level semantic information, namely object detection and semantic segmentation results. We will use both systems which are trained on the PASCAL VOC dataset which contains twenty classes, including humans, animals, vehicles, and indoor objects. We propose several object detection features which are derived from the detection of bounding boxes and the object proposals. Similarly, we derive semantic segmentation features by comparing the semantic segmentation results with the object proposals (see Sect. 4).

Before introducing the high-level features, we derive from object detection and semantic segmentation results we shortly describe the standard shape features which are directly computed from the object proposals. We will apply these features in our baseline method and in combination with the semantic features.

Shape Features

We extract 17 object proposal features, namely *shape features* which are based on the shape of the binary mask and its position in the image. The features described here for shape are existing features from the saliency literature. For every object proposal, we compute a set of shape features similar to the ones proposed in [24, 35]. The shape features we consider are centroid (2 dimensions), area, perimeter, convex area, Euler Number, major axis length, minor axis length, eccentricity, orientation, equivalent diameter, solidity, extent, width, and height of each object proposal. As an additional shape feature, we add the border-clutter feature [55] which is a binary feature indicating if the object proposal touches the boundaries of the image and is therefore cluttered by the field of view of the image. We also model the fact that salient objects are more frequent near the center of the image [23]. This feature is modeled by placing a Gaussian in the center of the image (for standard deviation $\sigma_x = \text{width}/4$ along the horizontal coordinates and $\sigma_y = \text{height}/4$ along the vertical coordinates was chosen). The centrality of object proposals is equal to the average value of the Gaussian over all pixels within the object proposal. It should be noted that this bias is a consequence of the fact how people take pictures, which are subsequently used in the datasets for saliency estimation. However, this bias is generally not present in images from many cameras; consider for example security cameras or robotics cameras. However, it should be noted that for datasets where such a bias does not exist this can be learned by the classifier we use, and thus, this feature would subsequently be ignored.

Fig. 3 Overview of feature extraction: we use an input image and a set of corresponding object proposals to compute the shape, object detection and semantic segmentation features



High-Level Semantic Features

The human visual system gives more attention to specific semantic object classes such as person, car, etc. In this section, we present high-level semantic features that we extract to compute saliency. These high-level features contain semantic knowledge of the object class. Therefore, the amount of saliency can depend on the semantic class which can be learned during the training phase.

Based on human perception, high-level features such as people, faces, and text have been proposed to capture visual attention [6, 9]. As for example [23] which assigns saliency to regions of faces, or the work of [6] which combines low-level bottom-up features with top-down features such as text. Other than these works, we consider a wider class of objects in this paper: the twenty classes of the PASCAL VOC includes person, animals, vehicles, and indoor objects. Recently, with deep learning the semantic understanding of images has improved significantly and currently is of high quality [20], we therefore think it is timely to evaluate the influence of a wider class of objects on saliency.

Object Detection Features (ODF)

Here, we propose several saliency features derived from object detection results. Object detectors in general detect a number of bounding boxes in the image. The detection provides a score related to an object class which indicates the confidence of the detector. Often, a threshold on the score is defined. Bounding boxes above this threshold are then considered detected objects.

In the pipeline which we described in Sect. 3, the aim is to assign saliency to object proposals. Therefore, to exploit high-level object detection we have to combine the object detection bounding boxes with the object proposals. To do so, we consider three different features which are all based on the intersection between detection bounding box and object proposals. They differ in the way they are normalized.

As a first measure, we consider the popular intersection over union, which is equal to the intersection of the i-th object proposal O_i and the j-th detection bounding box B_j divided by the union between the object proposal O_i and the detection bounding box B_j:

$$\mathrm{ODF}_1 = \frac{\left|O_i \cap B_j\right|}{\left|O_i \cup B_j\right|}, \tag{1}$$

where $\left|O_i \cap B_j\right|$ is equal to the number of pixels in set $O_i \cap B_j$. This measure is typically used in the evaluation of semantic segmentation [14].

The second measure computes the intersection over the minimum of the detection bounding box B_j and the object proposal O_i:

$$\mathrm{ODF}_2 = \frac{\left|O_i \cap B_j\right|}{\min(O_i,\ B_j)}, \tag{2}$$

and is sometimes considered as an alternative for intersection over union [48].

A drawback of the first measure is that in case the object proposal is part of the bounding box, but a significant part of the bounding box is outside the object proposal, this measure will assign a low saliency. The second measure addresses this problem; however, when the bounding box is included in the object proposal, this measure will assign a high saliency to the whole object proposal, even though the bounding box might only be a small part of the object proposal. Both these problems are addressed by the third measure which computes the percentage of pixels in object proposal O_i which are in the detection bounding box B_j:

$$\mathrm{ODF}_3 = \frac{\left|O_i \cap B_j\right|}{\left|O_i\right|}. \tag{3}$$

An example of the object detection features computation is shown in Fig 4. Comparison of object detection features computed on three example images (top row) from an example object proposal and object detection bounding box (bottom row). Superposed on the images in the bottom row are the object detection features. In these three examples, the saliency which should be assigned to the object proposal is high for the first two images and low for the last example. Only the third object detection feature correctly correlates with this.

It should be noted that we compute Eqs. 1–3 with the object proposal mask and with the bounding box representation for the detection. One could also decide to represent the object proposal with a bounding box, by drawing the smallest enclosing bounding box around the object proposal. Again, the same three features could be computed but now based on the bounding box for O_i. We compared both approaches on the PASCAL-S dataset and report the F-score (see Eq. 5) in Table 1. One can observe that using the original object proposal obtains better results than using bounding boxes. In addition, we see that the best results are obtained with object detection feature ODF_3. In all our experiments, we combine the three measures based on segmentation masks into the final ODF feature.

Semantic Segmentation Features (SSF)

As the second feature for high-level information, we use semantic segmentation results. Semantic segmentation

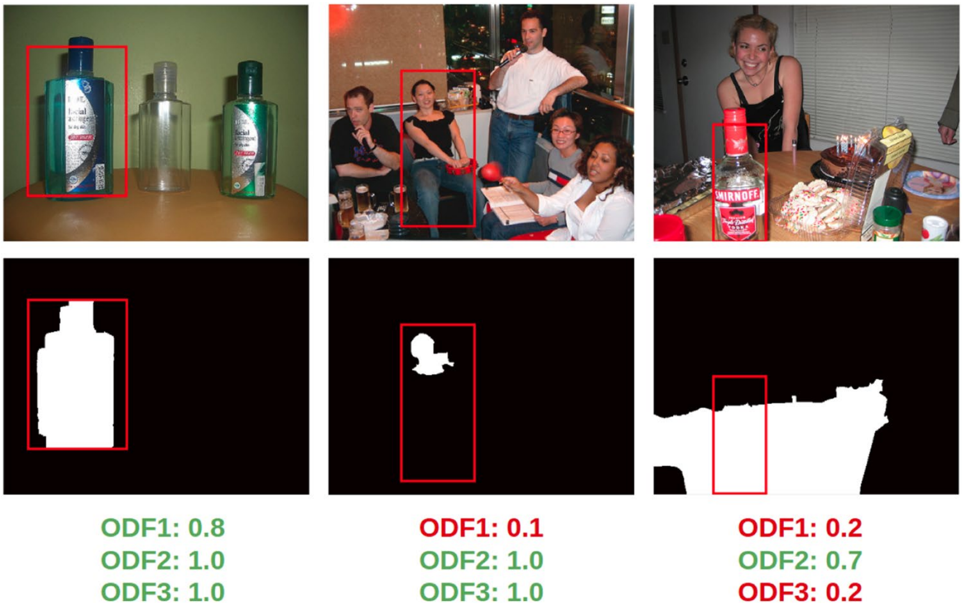


Fig. 4 Example of object feature computation for three example images. See text for details

Table 1 Comparison of detection features on the segmentation mask and the bounding box representation in terms of F-score

Features	ODF_1	ODF_2	ODF_3
Segmentation mask	45.40	58.90	64.30
Bounding box	45.10	30.080	53.30

algorithms output a probability map of the same size as the input image. For each pixel, it provides the probability that it belongs to one of the semantic classes. Typically, a background class is introduced for all pixels which do not belong to any of the semantic classes. Semantic segmentation can be considered a more difficult task than object detection because for good results the exact borders of objects need to be correctly detected.

We use the semantic segmentation results to propose a semantic segmentation feature (SSF) for saliency. For every semantic class c and object proposal O_i, we compute the SSF according to:

$$\mathrm{SSF}(c) = p\left(c|O_i\right) = \frac{\sum_{x\in O_i} p(c|x)}{|O_i|}, \tag{4}$$

where $p(c|x)$ is the output of the semantic segmentation algorithm and provides the probability of a semantic class conditioned on the pixel location x. The summation is over all pixels x included in the object proposal O_i. In this paper, we will evaluate semantic segmentation features derived from algorithms trained on PASCAL VOC, which has 21 classes, and therefore, the SSF feature of each object proposal will also have a dimensionality of 21.

Experiments and Results

In this section, we provide the implementation details, the experimental setup that we use in our approach, the benchmark datasets and the evaluation metrics.

Implementation Details

The overall pipeline of our method is provided in Sect. 3. Here, we report the implementation details.

Object Proposals Generation

From the input images, we compute a set of object proposals using the multiscale combinatorial grouping (MCG) method [1]. This method is based on a bottom-up hierarchical image segmentation. It was found to obtain improved results compared to other object proposal methods [25, 54]. We use the algorithm with default settings, which generates an average of 5153 object proposals per image.

Object Detection

To generate the object detection bounding boxes, we use fast R-CNN [20]. Fast R-CNN is an improved version of the R-CNN [19]; it obtains a significant speed-up by sharing the computation of the deep features between the bounding boxes. We use the fast R-CNN detector [20] which is trained on PASCAL VOC 2007.

Segmentation Results

For semantic segmentation, we use the approach proposed by Long et al. [33]. They compute the segmentation maps with a fully convolutional neural network (FCN) using end-to-end training. They improve the accuracy of their approach by using features extracted at multiple scales and adding skip connections between layers. We used the code provided by [33] and trained on the 20 classes of the PASCAL VOC 2011.

Random Forest

To assign saliency to every object, we use random forest and we set the number of trees to 200.

For each object proposal, we extract the feature vector of the SF, the ODF and the SSF features which are combined in a single feature vector. The random forest is trained on the training set. The ground-truth saliency score of object proposals is given by the ratio of pixels in the object proposal which are considered salient by the ground-truth divided by the total number of pixels in the object proposal. At testing time, we assign the saliency for all the object proposals by applying the random forest regressor. To compute the saliency based only on SF features, we train only on the features based on the shape features SF.

Saliency Features

We will compare results of several different saliency features. As a baseline, we will only use the shape features (SF) explained in Sect. 3. *ODF* refers to the method which is only based on object detection features and *SSF* refers to the method which only uses semantic segmentation features. Combinations of features are indicated as, e.g., *SF&ODF* for joining shape feature and object detection features.

Experimental Setup

Datasets

To evaluate the performance of the proposed method, we provide both qualitative and quantitative results on seven benchmark datasets: FT [2], ImgSal [34], ECSSD [57], PASCAL-S [35], MSRA-B [39], HKU-IS [37], and SOD [41]. The FT dataset contains 1000 images, most of which have one salient object. It provides the salient object ground truth which is provided by [60]. The ground truth in [60] is obtained using user-drawn rectangles around salient objects. The ImgSal dataset contains 235 images collected from the internet. It provides both fixations as well as salient object masks. The ECSSD dataset contains 1000 images. It is obtained by collecting images from the internet and PASCAL VOC, and the ground truth masks are labeled by five subjects. The PASCAL-S dataset contains 850 images and was built on the validation set of PASCAL VOC which has 20 classes of objects. MSRA-B dataset contains 5000 images which mostly contain a single salient object. It is one of the most used datasets for visual saliency estimation. The HKU-IS dataset contains 4447 images with pixelwise annotation of salient objects. The SOD dataset contains 300 images, which contain multiple salient objects per image either with low contrast or overlapping with the image boundary. In contrast to the other datasets, it often contains more than one salient object. All the datasets contain manually labeled ground truth.

When available, we use the predefined split into train and test set. On MSRA-B, there are 3000 training and 2000 testing images. On HKU-IS, there is a training set of 3000 images and a testing set of 1447 images. On SOD dataset, there are 100 images for training and 200 images for testing. On the other four datasets, we use 40% for training and 60% for testing.

Evaluation Metrics

We evaluate the performance of our method using F-measure and Precision-Recall curve (PR). The PR curves are computed by binarizing the saliency map at different thresholds and comparing it to the ground truth mask. The F-measure is defined as :

$$F_\beta = \frac{\left(1+\beta^2\right) \cdot \text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}, \tag{5}$$

where β^2 is set to 0.3 which is commonly used in the visual saliency literature [31, 35]. We use the maximum F-measure (max F_β), which was suggested in [42] as a good summary of the PR curve.

We conduct a qualitative and quantitative comparison of our method against the following existing methods: a context aware method (GOF) [18], deep saliency (DS) [30], discriminative regional feature integration (DRFI) [24], frequency tuned saliency (FT) [2], graph-based manifold ranking (GBMR) [58], local and global estimation (LEGS) [62], hierarchical saliency (HS) [57], multiscale deep features (MDF) [37], regional principal color-based saliency detection (RPC) [36], principal component analysis saliency (PCAS) [45], textural distinctiveness (TD) [50], deeply supervised salient object (DSS) [28], deep contrast learning (DCL) [38], and deep hierarchical saliency (DHS) [40].

Results

We start by evaluating the additional gain obtained when adding object detection features (ODF) and semantic segmentation features (SSF). The results for the seven datasets are provided in Tables 2 and 3. When we look at the performance of ODF and SSF alone, we observe that semantic segmentation provides much better features for saliency detection than object detection. We think that this is caused by the fact that segmentation algorithms provide pixelwise results rather than bounding boxes, and therefore, the saliency feature computation for each object proposal is more accurate.

Next, we consider the absolute gain which is obtained by adding ODF and SSF features to our baseline method (indicated by SF). For both features, and on all seven datasets, the features provide a significant improvement. This clearly shows the importance of high-level semantic features for saliency assignment. Again the improvement is largest when adding features derived from semantic segmentation. The best results are obtained on SOD and PASCAL-S datasets where an absolute gain of over 11% is reported. For PASCAL-S, this is partially caused by the fact that the object detector and the semantic segmentation algorithm have been trained on the PASCAL VOC dataset. Therefore, these images always contain classes which are detected (or segmented) by these algorithms. Note, however, that none of the images used for training the object detector (or segmentation algorithm) is included in the PASCAL-S dataset. On the other datasets, especially on ImgSal, ECSSD, and HKU-IS also large improvements of ~7% are obtained. The absolute gain is 1.70% and 4.37% in MSRA-B and FT datasets, respectively.

Table 2 F-measure of baseline (SF) and object detection feature (ODF), their combination and the absolute gain obtained by adding semantic object detection features

Dataset	SF	ODF	SF and ODF	Gain ODF
FT	84.23	57.26	85.30	1.07
ImgSal	67.19	54.76	71.30	4.11
ECSSD	77.47	64.57	79.40	1.93
Pascal-S	70.40	66.84	73.32	2.92
MSRA-B	89.50	62.75	89.90	0.40
HKU-IS	76.60	61.40	78.10	1.50
SOD	58.02	48.62	59.51	1.49

Table 3 F-measure of baseline (SF) and semantic segmentation feature (SSF), their combination and the absolute gain obtained by adding semantic segmentation features

Dataset	SF	SSF	SF and SSF	Gain SSF
FT	84.23	85.84	88.60	4.37
ImgSal	67.19	73.47	74.90	7.71
ECSSD	77.47	82.21	84.60	7.13
Pascal-S	70.40	81.16	81.80	11.40
MSRA-B	89.50	91.04	91.20	1.70
HKU-IS	76.60	82.16	83.30	6.70
SOD	58.02	66.64	69.90	11.88

Next, we compare our method to state-of-the-art saliency detection methods. The results are provided in Fig. 5. Overall, we obtain state-of-the-art on three of the seven datasets. On the FT dataset, we clearly outperform all the other salient object detection methods with both object detection and segmentation features. Also, we obtain the best F-measure compared to other state-of-the-art methods. On the ImgSal dataset, we also achieve the best performance on F-measure. The performance is better over a wide range of recalls only to be slightly outperformed for the highest recalls by GOF.

On the ECSSD dataset, we obtain the best results when considering only those which are trained on the ECSSD training dataset, but we are outperformed by recent end-to-end trained networks trained on MSRA-B. Similar results are obtained on the PASCAL-S dataset. Only DSS and DHS outperform our semantic segmentation-based method. DS and MDF methods outperform our object detection-based method. On the MSRA-B dataset, we are outperformed by DSS and DCL methods which are recent end-to-end trained saliency methods but still obtain competitive results of 91.20. Similarly, on the HKU-IS dataset we obtain competitive results. Finally, on the SOD dataset we outperform all the other salient object detection methods

To get better insight in which classes contribute to the improvement in saliency detection, we have performed an additional experiment using SSF. We investigate which semantic classes are important. We perform this analysis on four datasets: FT, ImgSal, ECSSD, and PASCAL-S. We evaluate the drop of saliency if we remove one class, the results are shown in Fig 6. The results show that removing both bird and person significantly deteriorates saliency estimates on all four datasets. Some other classes contribute only on some of the datasets. Examples are aeroplane, bicycle, potted plant, sofa and tv-monitor which lead to a drop of over 0.6 when removed. Removing some classes actually leads in some cases to a small increase in performance, possibly caused to overfitting or noise in the semantic segmentation algorithm.

We provide a qualitative comparison in Fig. 7. We tested our method in several challenging cases, low contrast between object and background (first two rows), results of objects touching the image boundary are shown where our method successfully includes the regions that touch the border (third and fourth row). Finally, the case when multiple disconnected objects is investigated (last two rows).

Finally, as an indication of the computational complexity of the method, we have computed timings of our algorithm on the PASCAL-S dataset. It takes an average of 39.8s/

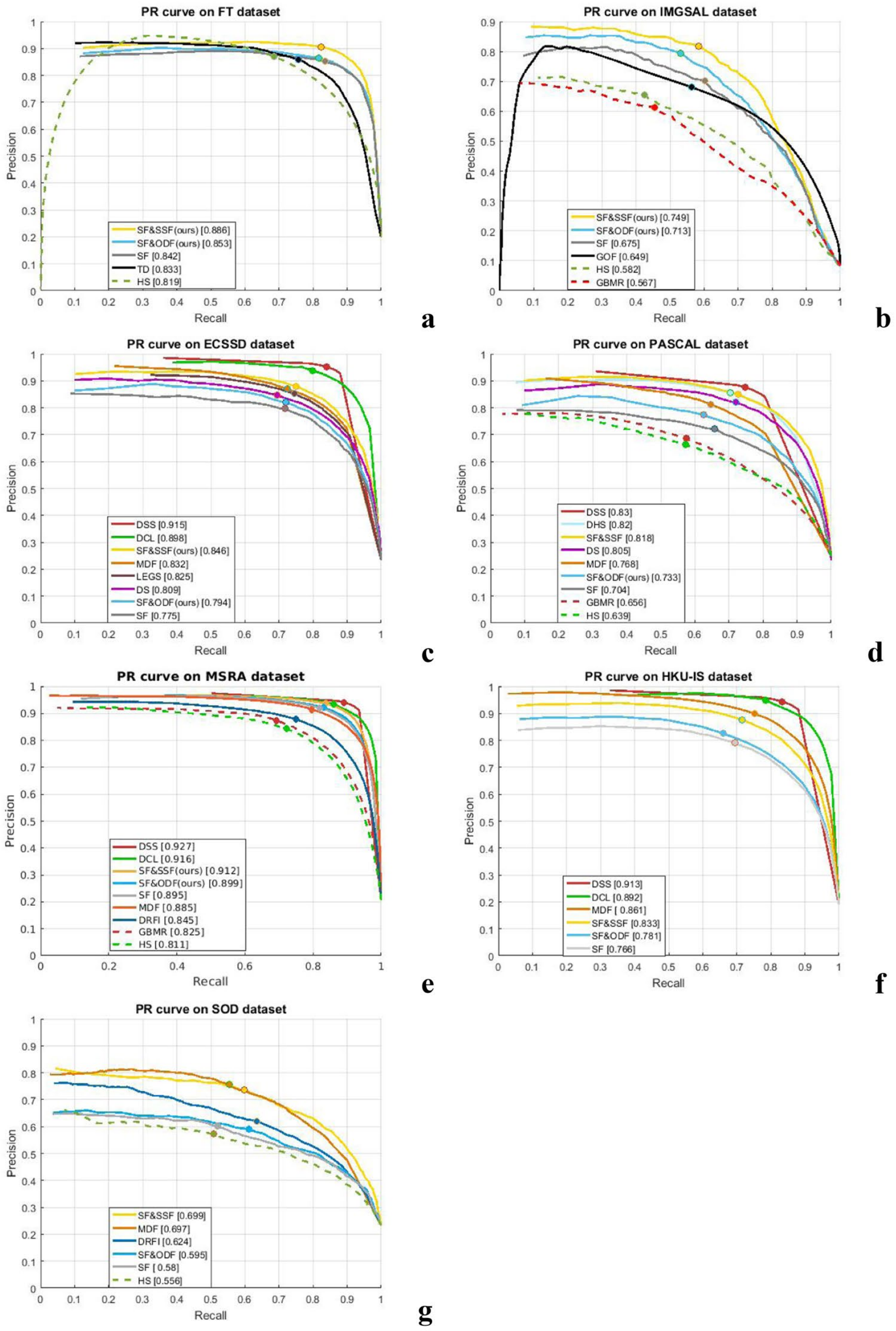


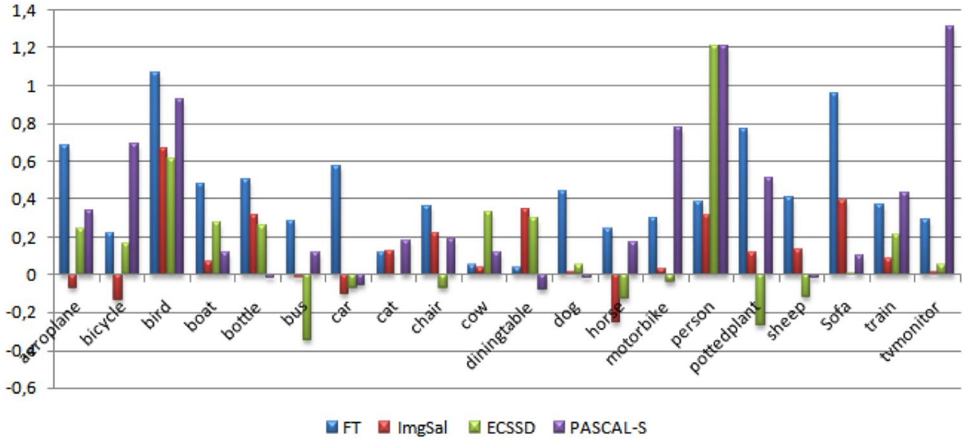
Fig. 5 PR curves for a variety of methods, on several datasets: **a** FT, **b** ImgSal, **c** ECSSD, **d** Pascal-S, **e** MSRA-B, **f** HKU-IS, and **g** SOD

image (computation based on 100 images). It should be noted that most of the time was spent by the MCG algorithm 35.3s, the extraction of the features took 3.4s and the classifier 1.1s. Furthermore, this is based on an unoptimized Matlab implementation.

Conclusion

The importance of high-level semantic image understanding on saliency estimation is known [9, 43]. However, most computational methods are bottom-up or only include few semantic classes such as faces and text [6, 23]. Therefore, we have evaluated the impact of recent advances in high-level

Fig. 6 Saliency drop (measured by the F-score) as a consequence of removing a single semantic class on the four datasets. From left to right: FT, ImgSal, ECSSD and Pascal-S



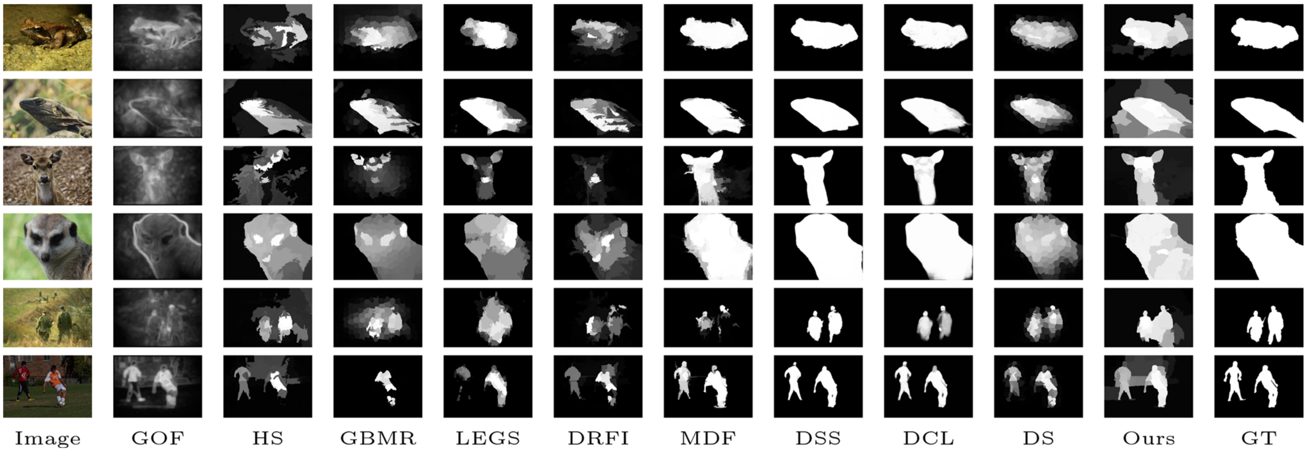


Fig. 7 Qualitative comparison of saliency maps generated from nine state-of-the-art methods. From left to right: input image, GOF [18], HS [57], GBMR [58], LEGS [62], DRFI [24], MDF [37], DSS [28], DCL [38], DS [30], Ours, and ground truth

semantic image understanding on saliency estimation. In order to do that, we have derived saliency features from two popular algorithms: fast-RCNN for object detection and FCNs for semantic segmentation. We found that the features based on semantic segmentation obtained superior results, most probably due to the fact that they provide pixel-wise labels, which lead to more accurate saliency estimation maps. To evaluate the derived features from object detection and semantic segmentation, we perform experiments on several standard benchmark datasets. We show that a considerable gain is obtained from the proposed features and we examine which semantic class boost more the task of saliency. We found that the classes of person and bird are among the most important.

Furthermore, in the evaluation on seven benchmark datasets we outperform state-of-the-art on three of them (FT, ImgSal, and SOD) and obtain competitive results on the other four (ECSSD, PASCAL-S, MSRA-B, and HKU-IS).

One of the limitations of the proposed approach is the computational time. However, the vast majority of this time is spent on object proposal extraction, where we use the MCG method. However, there is active research on fast object proposal methods which can reduce the current time of 35.3s to around 0.26s [29]. For future work, we will evaluate the usage of these fast object proposal methods within our framework. For further future work, we are interested in extending current end-to-end networks for saliency with explicit modules for object detection and evaluate if such architectures could further improve state-of-the-art approaches. It would also be interesting to evaluate the impact of a larger set of object classes on saliency detection (currently, we evaluate the 20 classes from the PASCAL VOC challenge). Finally, in this paper we evaluated the impact of high-level information on salient object detection, but it would also be interesting to perform a similar study for saliency maps derived from eye-tracking experiments.

Acknowledgements We acknowledge the Spanish project TIN2016-79717-R, the CHISTERA project M2CR (PCIN-2015-251) and the CERCA Programme/Generalitat de Catalunya. Marc Masana acknowledges 2018FI_B1_00198 Grant of *Generalitat de Catalunya*. We also acknowledge the generous GPU support from Nvidia.

Compliance with Ethical Standards

Conflict of interest Author Aymen Azaza declares that he has no conflict of interest. Author Joost van de Weijer declares that he has no conflict of interest. Author Ali Douik declares that he has no conflict of interest. Author Javad Zolfaghari declares that he has no conflict of interest. Author Marc Masana declares that he has no conflict of interest.

References

1. Arbelaez P, Pont-Tuset J, Barron J, et al. 'Multiscale combinatorial grouping'. IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014;328–335.
2. Achanta R, Hemami S, Estrada F, et al. 'Frequency-tuned salient region detection'. IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009;1597–1604.
3. Azaza A, van de Weijer J, Douik A, et al. Context proposals for saliency detection. Computer Vision and Image Understanding, Elsevier. 2018;174:1–11.
4. Bylinskii Z, Recasens A, Borji A, et al.: 'Where should saliency models look next?'. In European Conference on Computer Vision, Amsterdam, The Netherlands, 2016;809–824.
5. Bian P, Zhang L. Biological plausibility of spectral domain approach for spatiotemporal visual saliency. In International conference on neural information processing: Springer, Berlin Heidelberg; 2008. p. 251–8.
6. Borji A. 'Boosting bottom-up and top-down visual features for saliency estimation'. IEEE Conference In Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 2012;438–445.
7. Borji A, Cheng M-M, Jiang H, et al. 'Salient object detection: A survey', arXiv preprint 2014;arXiv:1411.5878.
8. Carreira J, Caseiro R, Batista J, et al. Semantic segmentation with second-order pooling. Computer Vision ECCV: Firenze, Italy; 2012. p. 430–43.
9. Cerf M, Frady EP, Koch C. Faces and text attract gaze independent of the task: Experimental data and computer model. Journal of vision. 2009;9(12):15–15.
10. Cheng MM, Zhang Z, Lin WY, et al. 'BING: Binarized normed gradients for objectness estimation at 300fps'. In Proceedings of the IEEE conference on computer vision and pattern recognition, Columbus, OH, USA, 2014;3286–3293.
11. Coutrot A, Guyader N. 'Learning a time-dependent master saliency map from eye-tracking data in videos', arXiv preprint 2017; arXiv:1702.00714
12. Ehinger KA, Hidalgo-Sotelo B, Torralba A, et al. Modelling search for people in 900 scenes: A combined source model of eye guidance. Visual cognition. 2009;17(6–7):945–78.
13. Einhauser W, Spain M, Perona P. Objects predict fixations better than early saliency. Journal of Vision. 2008;8(14):18–18.
14. Everingham M, Van Gool L, Williams CK, et al. The pascal visual object classes (voc) challenge. International journal of computer vision. 2010;88(2):303–38.
15. Fang Y, Chen Z, Lin W, et al. Saliency detection in the compressed domain for adaptive image retargeting. IEEE Transactions on Image Processing. 2012;21(9):3888–901.
16. Felzenszwalb PF, Girshick RB, McAllester D, et al. Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence. 2010;32(9):1627–45.
17. Frintrop S, Werner T, Martin Garcia G. 'Traditional saliency reloaded: A good old model in new shape'. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, USA, 2015;82–90.
18. Goferman S, Zelnik-Manor L, Tal A. Context-aware saliency detection. IEEE Transactions on Pattern Analysis and Machine Intelligence'. 2012;34(10):1915–26.
19. Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. Computer Vision and Pattern Recognition: Columbus, OH, USA; 2014. p. 580–7.
20. Girshick R. 'Fast R-CNN'. International Conference on Computer Vision (ICCV), Santiago, Chile, 2015;1440–1448.
21. Hou X, Zhang L. 'Saliency detection: A spectral residual approach'. IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 2007;1–8.
22. Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis & Machine Intelligence. 1998;20(11):1254–9.
23. Judd T, Ehinger K, Durand F, et al. 'Learning to predict where humans look'. 12th International Conference on Computer Vision, Kyoto, Japan , 2009;2106–2113.
24. Jiang H, Wang J, Yuan Z, et al. 'Salient object detection: A discriminative regional feature integration approach'. IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 2013;2083–2090.
25. Krähenbühl P, Koltun V. 'Geodesic object proposals'. Springer, European Conference on Computer Vision, Zurich, Switzerland, 2014;725–739.
26. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012;1097–105.
27. Kalboussi R, Azaza A, van de Weijer J, et al. Object proposals for salient object segmentation in videos. Multimedia Tools and Applications. 2019;1–17.
28. Hou Q, Cheng MM, Hu X, Borji A, et al. 'Deeply supervised salient object detection with short connections'. In: CVPR. 2017;5300–5309.
29. Hu H, Lan S, Jiang Y, Cao Z, Sha F. Fastmask: Segment multiscale object candidates in one shot. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017;991–9.
30. Xi L, Liming Z, Lina W, et al. DeepSaliency: Multi-Task Deep Neural Network Model for Salient Object Detection. IEEE Transactions on Image Processing. 2016;25(8):3919–30.
31. Lin H, Li J, Liang DC, et al. Saliency detection using adaptive background template. IET Computer Vision. 2017;11(6):389–97.
32. LeCun Y, Boser B, Denker JS, et al. Backpropagation applied to handwritten zip code recognition. Neural computation. 1989;1(4):541–51.
33. Long J, Shelhamer E, Darrell T. 'Fully convolutional networks for semantic segmentation'. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015;3431–3440.

34. Li J, Levine MD, An X, et al. Visual saliency based on scale-space analysis in the frequency domain. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2013;35(4):996–1010.
35. Li Y, Hou X, Koch C, et al. 'The secrets of salient object segmentation'. IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014;280–287.
36. Lou J, Ren M, Wang H. 'Regional Principal Color Based Saliency Detection', PLoS ONE, 2014;9:(11).
37. Li G, Yu Y. 'Visual Saliency Based on Multiscale Deep Features'. IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015;5455–5463.
38. Li G, Yu Y. 'Deep Contrast Learning for Salient Object Detection'. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 2016;478–487.
39. Liu T, Yuan Z, Sun J, et al. Learning to detect a salient object. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2011;33(11):353–67.
40. Liu N, Han J. ' Dhsnet: deep hierarchical saliency network for salient object detection'.Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016;678–686.
41. Martin D, Fowlkes C, Tal D, Malik J. 'A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics,' in Proc. IEEE Conf. ICCV, 2001;416–423.
42. Martin DR, Fowlkes CC, Malik J. Learning to detect natural image boundaries using local brightness, color, and texture cues. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2004;26(5):530–49.
43. Mackworth NH, Morandi AJ. The gaze selects informative details within pictures. Springer, Attention, Perception, & Psychophysics. 1967;2(11):547–52.
44. Jiang Ming, Boix Xavier, Xu Juan, Roig Gemma, Van Gool Luc, Zhao Qi.' Saliency Prediction with Active Semantic Segmentation', In Xianghua Xie, Mark W. Jones, and Gary K. L. Tam, editors, Proceedings of the British Machine Vision Conference (BMVC) 2015;15.1-15.13.
45. Margolin R, Tal A, Zelnik-Manor L. 'What makes a patch distinct?'. IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 2013;1139–1146.
46. Nuthmann A, Henderson JM. Object-based attentional selection in scene viewing. Journal of vision. 2010;10(8):20–20.
47. Perazzi F, Krähenbühl P, Pritch Y, et al. 'Saliency filters: Contrast based filtering for salient region detection'. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 2012;733–740
48. Senior, A., Senior, A. W.: 'Protecting privacy in video surveillance', (New York: Springer.ISO 690, 1st edn.), pp. 11–33 (2009)
49. Razavian S, Azizpour A, Sullivan H *et al.*: 'CNN features off-the-shelf: an astounding baseline for recognition'. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 2014;806–813
50. Scharfenberger C, Wong A, Fergani K et al. 'Statistical textural distinctiveness for salient region detection in natural images'. IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 2013;979–986.
51. Stella X, Yu, Lisin DA. 'Image compression based on visual saliency at individual scales', Springer, Advances in Visual Computing, 2009;157–166.
52. Treisman AM, Gelade G. A feature-integration theory of attention. Elsevier, Cognitive psychology. 1980;12(1):97–136.
53. Tingtian Li, Lun Daniel PK. 'A Novel Reflection Removal Algorithm Using the Light Field Camera', IEEE International Symposium on Circuits and Systems (ISCAS), 2018;1–5.
54. Uijlings JR, Van De Sande KE, Gevers T, et al. Selective search for object recognition. International journal of computer vision. 2013;104(2):154–71.
55. Qin Y, Lu H, Xu Y et al. 'Saliency Detection via Cellular Automata'. IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015;110–119.
56. Xu J, Jiang M, Wang S, et al. Predicting human gaze beyond pixels. Journal of vision. 2014;14(1):28–28.
57. Yan Q, Xu L, Shi J et al. 'Hierarchical saliency detection'. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 2013;1155–1162.
58. Yang C, Zhang L, Lu H et al. 'Saliency detection via graph-based manifold ranking'. IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 2013;3166–3173.
59. Yang J, Yang MH. Top-down visual saliency via joint crf and dictionary learning. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2016;39(3):576–88.
60. Wang Z, Li B. 'A two-stage approach to saliency detection in images'. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', Las Vegas, NV, USA, 2008;965–968
61. Wan S, Jin P, Yue L. 'An approach for image retrieval based on visual saliency'. International Conference on Image Analysis and Signal Processing, Linhai, China, 2009;172–175.
62. Wang L, Lu H, Ruan X et al. 'Deep Networks for Saliency Detection via Local Estimation and Global Search'. IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015;3183–3192.
63. Wang G, Zhang Y, Li J. High-level background prior based salient object detection'. Journal of Visual Communication and Image Representation. 2017;48:432–41.
64. Zitnick CL, Dollar P. 'Edge boxes: Locating object proposals from edges'. In European Conference on Computer Vision, Zurich, Switzerland, 2014;391–405.
65. Zhu C, Li G, Wang W et al. 'Salient Object Detection with Complex Scene based on Cognitive Neuroscience'. In : Multimedia Big Data (BigMM), 2017 IEEE Third International Conference on IEEE, California, USA, 2017;33–37.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.