Image and Vision Computing 31 (2013) 31–42

Contents lists available at SciVerse ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Salient object detection: From pixels to segments ☆

Victoria Yanulevskaya [a,*], Jasper Uijlings [a], Jan-Mark Geusebroek [b]

[a] Department of Information Engineering and Computer Science, University of Trento, Trento, Italy
[b] Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 13 January 2011
Received in revised form 5 June 2012
Accepted 21 September 2012

Keywords:
Salient object detection
Object-based visual attention theory
Proto-objects

ABSTRACT

In this paper we propose a novel approach to the task of salient object detection. In contrast to previous salient object detectors that are based on a spotlight attention theory, we follow an object-based attention theory and incorporate the notion of an object directly into our saliency measurements. Particularly, we consider proto-objects as units of the analysis, where a proto-object is a connected image region that can be converted into a plausible object or object-part, once a focus of attention reaches it. As the object-based attention theory suggests, we start with segmenting a complex image into proto-objects and then assess saliency for each proto-object. The most salient proto-object is considered as being a salient object.
We distinguish two types of object saliency. Firstly, an object is salient if it differs from its surrounding, which we call center-surround saliency. Secondly, an object is salient if it contains rare or outstanding details, which we measure by integrated saliency. We demonstrate that these two types of object saliency have complementary characteristics; moreover, the combination of the two performs at the level of state-of-the-art in salient object detection.

1. Introduction

Many people will easily and consistently point at the salient object in the images presented in Fig. 1. Indeed, these images were carefully preselected with a two-stage labelling process to ensure a salient object is standing out from the background [1]. Nevertheless, salient object detection is still a challenging task for computer vision algorithms.

There are two prominent theories for human visual attention: spatial-based and object-based attention. Within the spatial-based theory, attention is compared to a spotlight or a zoom lens, which shifts our focus from one spatial location to another to sample the surrounding. As a result, all visual content within a fovea-sized region around these locations is processed [2,3]. As Fig. 2 demonstrates, such region can contain an object, parts of different objects, or parts of an object and its background. In contrast, the object-based attention theory argues that attention is actually focused on objects or so called proto-objects, for a review see [4]. A proto-object is defined as an unit of visual information that can be converted into a plausible object or object-part, once a focus of attention reaches it [5,6]. Fig. 3 sketches a way the object-based attention could work. The theory implies that at the early pre-attentive stage, the visual system pre-segments a complex scene into proto-objects. Then our focus is shifted from one proto-object to another. In this paper, we propose a novel salient object detector based on the object-based attention theory.

Salient object detection implies localization of a complete object of any class which attracts attention. In the literature [1,7,8] the standard way to approach this problem is to calculate a saliency map at the pixel level and then detect an image area which maximizes

Fig. 1. Example of images with a pronounced salient object taken from the dataset [1]. Rectangles represent human annotation.

☆ This paper has been recommended for acceptance by Ioannis Patras.

* Corresponding author. Tel.: +31 626409078.
E-mail addresses: yanulevskaya@disi.unitn.it (V. Yanulevskaya), jrr.uijlings@disi.unitn.it (J. Uijlings), J.M.Geusebroek@uva.nl (J.-M. Geusebroek).

0262-8856/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.imavis.2012.09.009

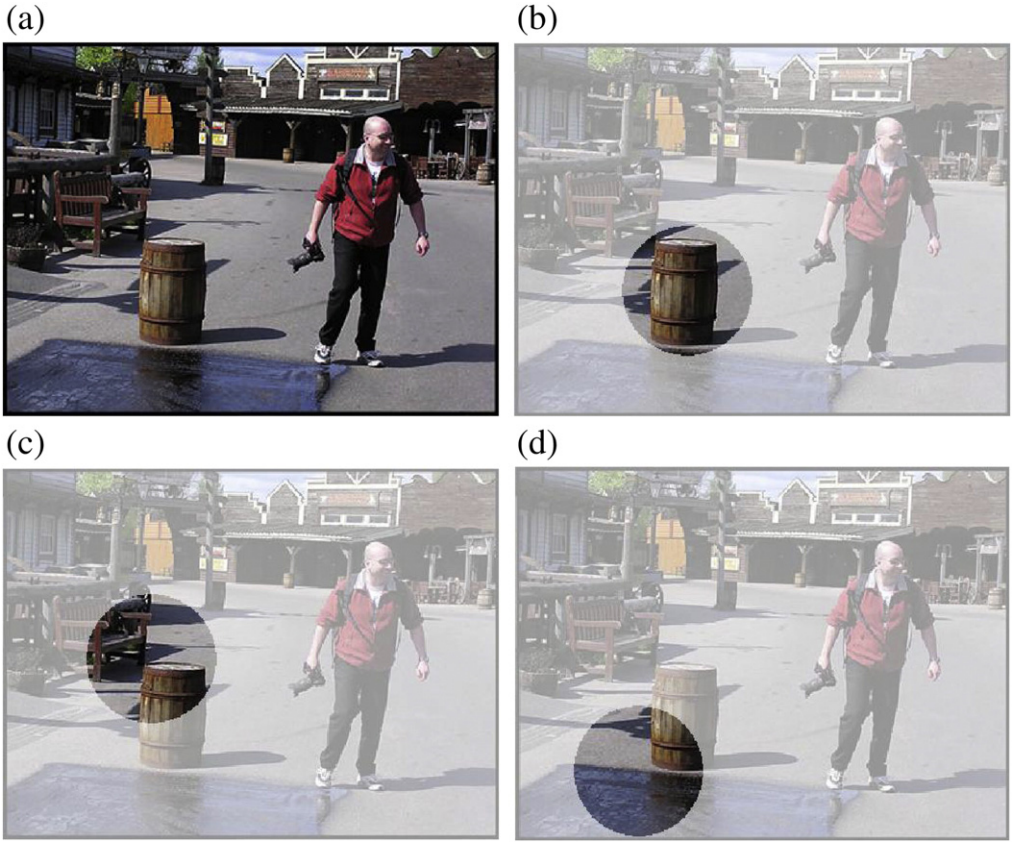

Fig. 2. An illustration of a spatial-based theory for visual attention. High contrast regions represent a focus of attention. (a) An input image. Focus of attention contains (b) a single object, (c) parts of different objects, (d) a part of an object and its background.

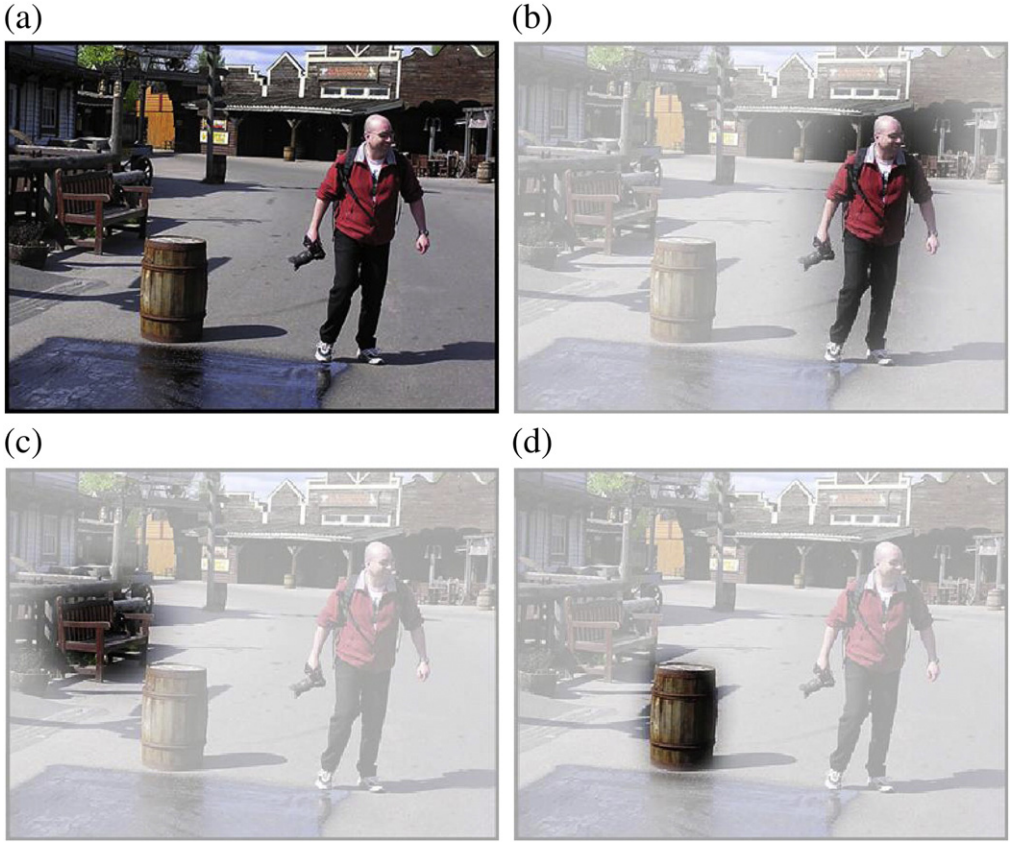

Fig. 3. An illustration of an object-based theory for visual attention. High contrast regions represent a focus of attention. (a) An input image. Focus of attention always contains an object: (b) a person, (c) a bench, and (d) a barrel.

saliency based on some localization technique such as sliding windows. Thus, a salient object is determined mostly by the structure of the saliency map. However, such map does not take into account explicitly the information about objects of an image. Therefore, most of the recent salient object detectors follow the spatial-based attention theory.

In this work, we propose to incorporate the notion of an object directly into the saliency measurements. As the object-based attention theory suggests we start with the segmentation of a complex image into proto-objects. Although general object segmentation is a hard task, roughly outlining important segments in an image by feature grouping is certainly doable. Then we assess saliency for each proto-object and report the most salient one.

Fig. 4 illustrates the advantage of the proposed object-based method in comparison with the spatial-based approach. Spatial-based approaches look for the most salient spot in an image. As a result, it might mix parts of different objects or detect only prominent object details. As it is shown in Fig. 4, the most salient image window (c) selects a region which contains the most outstanding detail of the bollard together with trees, whereas both windows (d) and (e) capture only prominent but small parts of the bollard. In contrast, our method is object-based, and therefore, assesses saliency for connected image regions. Thus, it succeeds to separate different salient objects. In Fig. 4(f) the most salient image segment contains only the bollard, which is correct, while trees are in the separate segment (h). Furthermore, although the bollard has only few outstanding details, the object-based approach encourages detection of the complete object (f), or its upper part (g), which might be considered as an object itself as the rope clearly divides the bollard in two parts. This indicates that the object-based theory for attention might be better suited for salient object detection.

We estimate object saliency in two ways. First, we measure how an object as a whole differs from its background. We call this center-surround saliency. Second, we calculate summed rarity of details within an object. We call this integrated saliency. We combine both types of saliency as they have complementary characteristics.

In this work, we propose a salient object detector which follows the object-based attention theory. The novelty of the proposed method is in the way the proto-object based theory for visual attention is applied for salient object detection. The standard approaches [1,6–8] can be divided in two steps. During the first step, the saliency is estimated at pixel level based on some local features. Afterwards, at the second step, the saliency map is used to obtain object hypotheses. Thus, in traditional approaches, proto-objects are determined mostly by the structure of the pixel level saliency map whereas the information about image objects is not taken into account explicitly. In contrast, we start with proto-object extraction by segmenting the image itself, rather than segmenting its saliency map. Afterwards, we estimate the saliency directly at the proto-object level. It allows for incorporating the notion of an object into saliency measurements by considering a proto-object as a unit of analysis. We demonstrate that the proposed method achieves state-of-the-art performance on a standard dataset [1].

2. Related work

In the task of salient object detection it is common to follow a spatial-based attention theory by calculation a pixel-based saliency map with consecutive localization of the image area which maximizes saliency. Liu *et al.* [1] consider multi-scale contrast, center-surround colour histograms, and colour spatial distribution to calculate pixel saliency. At the localization step, all features are combined in a conditional random field resulting in a binary label map which separates the salient object from the background. This method demonstrates a good performance. However, it involves learning the CRF which in general is computationally expensive. Valenti *et al.* [8] present a real time salient object detector with similar performance to [1]. In their method, pixel saliency is calculated as a linear combination of three features:

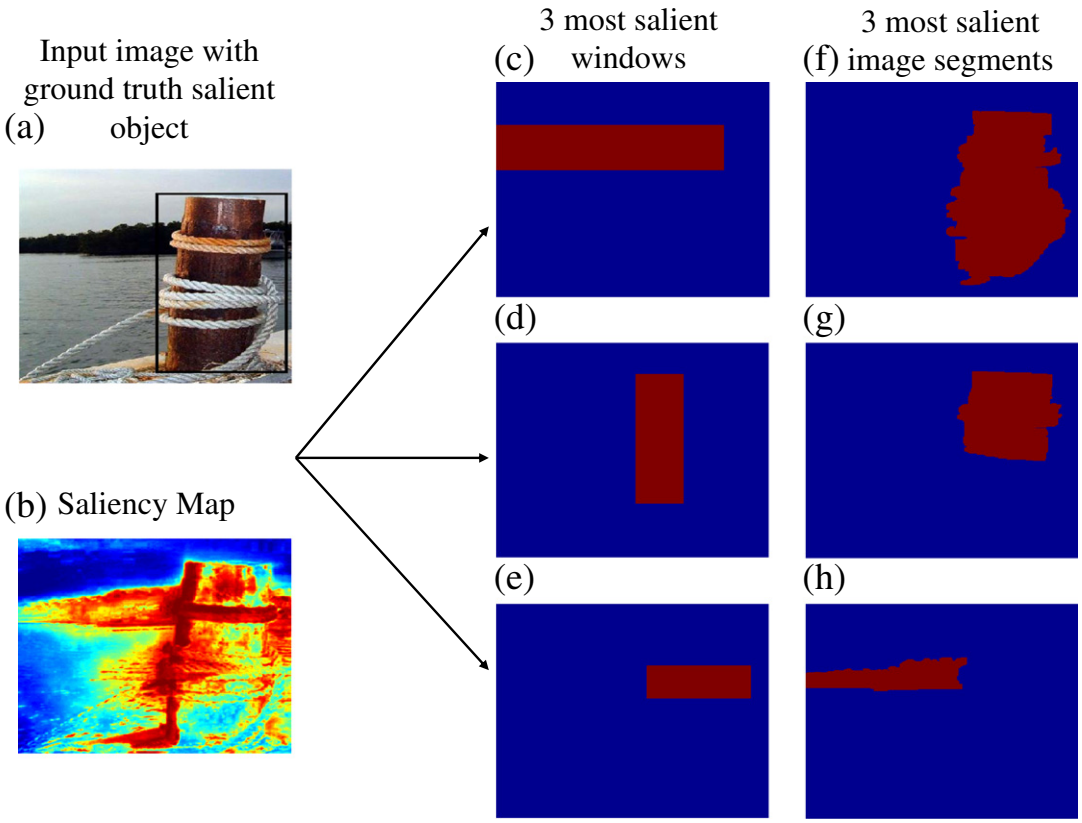

Fig. 4. A spatial-based versus object-based approach for salient object detection. (a) Input image. (b) Pixel-based saliency map. (c)–(e) Most salient windows of an image according to a spatial-based approach. (f)–(h) Most salient proto-objects of an image according to an object-based approach.

isocentricity, curvedness, and rarity of colour edges. This approach highlights centers and edges of the image structures. In order to distribute saliency within connected regions, the authors run a graph-based segmentation and average values of the saliency map inside each segment. At the localization step, efficient subwindow search [9] is used. In contrast, our method follows an object-based attention theory and calculates saliency of segments of an image. Thus, we automatically distribute saliency within connected regions. More importantly, we incorporate a notion of an object into our saliency measure.

Marchesotti *et al.* [7] propose a salient object detector which is based on the assumption that images with similar appearance are likely to have salient objects with the same characteristics. To measure saliency within a target image, the authors train a classifier on the *K* most similar images, with provided ground truth bounding boxes around salient objects. Two-class classification problem is considered: the salient class consists of salient objects, and the non-salient class consists of the background. Each patch of the target image is classified as being salient/non-salient. In order to locate a salient object, the output of the classifier is used to initialize an iterative graph-cut algorithm inspired by [10]. As a result, the segment which covers most of the salient pixels is reported as the salient object. The method is shown to achieve very promising results when annotated image data is available. However, the authors have also shown that the method is highly dependent on the quality of the retrieval step in which the most similar images are extracted. In contrast, our method does not rely on any learning; hence, it does not require image annotation and retrieval.

Walther and Koch [6] propose a way to extract proto-objects based on the spatial-based approach by Itti *et al.* [11]. The spatial-based model by Itti *et al.* [11] results in a pixel-based saliency map. Walther and Koch [6] define proto-objects as spatial extension of the peaks of this saliency map. In fact, an extracted proto-object consists of a set of pixels which is defined by a continuous 4-connected neighborhood of a peak with saliency above a certain threshold. Therefore, in the Walther and Koch's approach, the most salient points are calculated according to the spatial-based model, afterwards the saliency is spread to the region around them. Hence this means that their proto-objects are extracted from the saliency map. In contrast, we fully rely on the object-based attention theory and extract proto-objects directly from the image by feature grouping. This allows us to assess saliency at the proto-object level.

We extract proto-objects by dividing an image into coherent regions which are feasible candidates for salient objects. To do this, we follow the strategy introduced by [12], who adapt hierarchical segmentation to find good candidates for object locations. The work by [12] is based on two key ideas. First of all, objects can be of any size and can occur at any scale. Therefore a hierarchical segmentation strategy is used and all segments throughout the whole hierarchy are considered. Second, in order to account for different object appearances and image conditions, the results of several, complementary segmentations are combined. This strategy has proven itself successful in an object localization task [12]. To the best of our knowledge, we are the first to apply it for the task of salient object detection.

To establish integrated saliency, we follow the information maximization approach [13] to measure rarity of object details. Intuitively, image locations which deviate from the rest of an image should be salient. Bruce and Tsotsos [13] define saliency based on maximum information sampling. They calculate the Shannon's self-information based on the likelihood of the local image content in a patch given

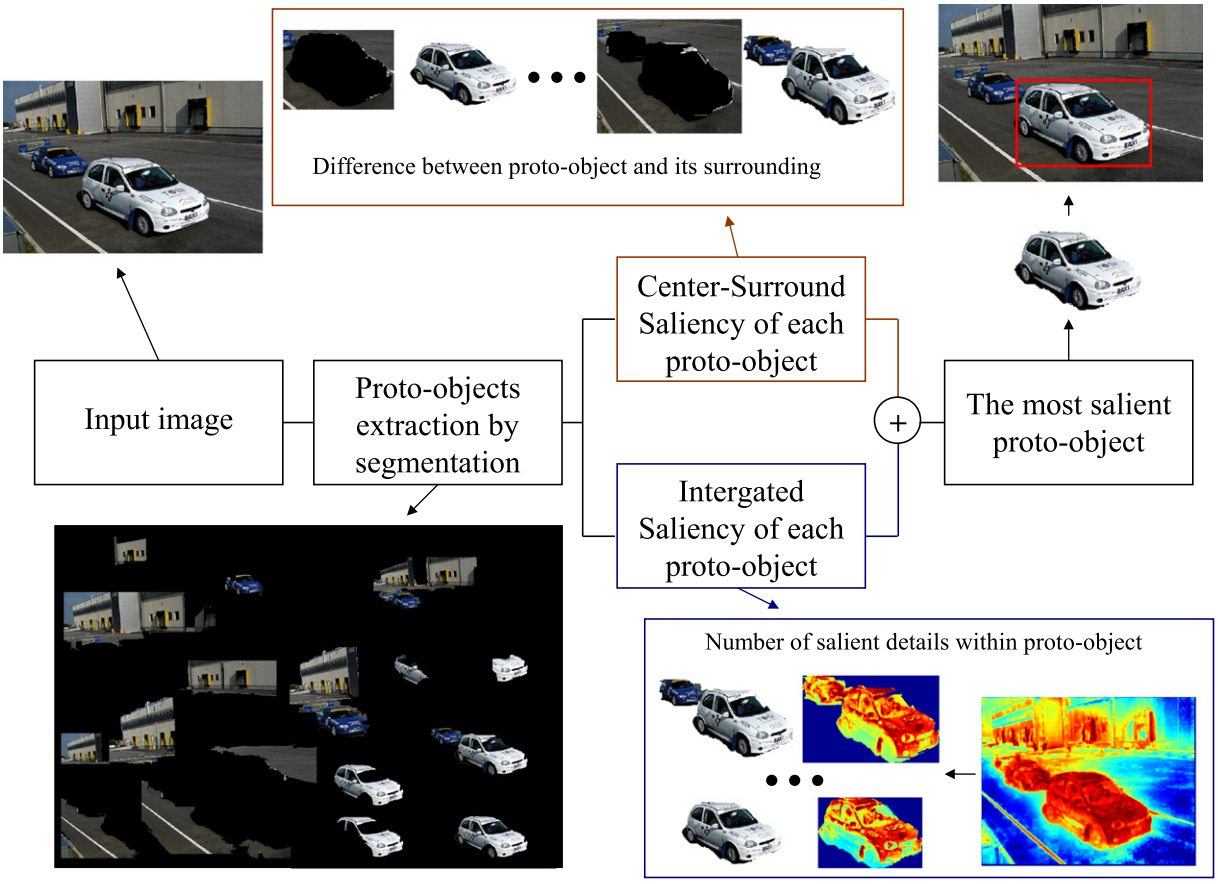

Fig. 5. Overview of the proposed salient object detector. Given an input image our aim is to find the most salient object. We start with a hierarchical segmentation to generate many candidate proto-objects for a salient object. We assess the saliency of all segments: Firstly, we estimate how much the entire segment pops out from its surroundings (center-surround saliency). Secondly, we measure how many details are within a segment which pop out with respect to the entire image (integrated saliency). We combine both types of saliency and select the segment with the highest value.

the content of the rest of the image. Patches with unexpected content are more informative, and thus salient. To reflect image content we have chosen to use visual words and colour histograms. We will demonstrate that saliency based on these features outperforms traditional information maximization saliency [13] and standard spectral residual saliency [14] for the task of salient object detection.

3. Methods

An overview of the proposed method is presented in Fig. 5. In our analysis, we follow the object-based attention theory. This theory assumes that attention focuses on proto-objects, which are plausible candidates for salient objects. As we do not assume any prior knowledge about a salient object such as its type, colour, or size, we use a set of hierarchical image segmentations to obtain a high variety of proto-objects [12]. Although not all segments are perfect candidates for real-life objects, there is a high probability that within this set some segments accurately separate objects from the surrounding. As can be seen in Fig. 5, sometimes only a front side or a door of the white car are extracted; however, there is also a proto-object which outlines the complete white car carefully. To find the best candidate for a salient object we measure the saliency of each proto-object in two ways. (1) With center-surround saliency we measure how the proto-object differs from its surrounding in terms of colour histogram. (2) We calculate the integrated saliency by measuring the energy of a saliency map within a proto-object. In the remaining of this section we provide more details of each part of our method.

3.1. Hierarchical image segmentation

We adapt the novel approach by [12] to obtain a set of candidate proto-objects. As a starting point we over-segment an image using the publicly available code of Felzenszwalb and Huttenlocher [15]. We then follow a standard grouping procedure where each segment is represented by a vertex and neighbouring pairs are represented by edges. For each edge we calculate a similarity between segments based on four characteristics. Like [12], we use texture distribution and size of segments. Additionally, we consider colour distribution and spatial relationship, where we found the later to be particularly helpful (data is not shown). Then we iteratively select the edge with the highest similarity, merge the corresponding segments, and calculate all similarities with this new segment and its neighbours. We repeat this until the whole image becomes a single segment. Our similarity function $S(a,b)$ consists of four components in range [0,1] and is defined as:

$$S(a,b) = S_{\text{colour}}(a,b) + S_{\text{texture}}(a,b) + S_{\text{enclosed}}(a,b) + S_{\text{size}}(a,b). \quad (1)$$

The colour-based similarity between segments $S_{\text{colour}}(a,b)$ is calculated as a histogram intersection between their opponent colour histograms [16]. We use histogram intersection instead of the χ^2 distance for the sake of computational efficiency. The texture based similarity $S_{\text{texture}}(a,b)$ is calculated as a histogram intersection of gradient histograms of segments in horizontal and vertical directions in opponent colour space.

Similarity $S_{\text{enclosed}}(a,b)$ reflects the spatial relationship between segments. Let $Bn(a)$ be defined as the number of boundary pixels of a and $Bn(a)<Bn(b)$, then

$$S_{\text{enclosed}}(a,b) = \frac{Br(a,b)}{Bn(a)}, \quad (2)$$

where $Br(a,b)$ counts the number of pixels of segment a that touch segment b. If a is completely enclosed by b (i.e. a fills a hole in b), then $S_{\text{enclosed}}(a,b)$ is one. If a touches b with only a single pixel, it is near zero. With this component we encourage to fill holes inside a segment.

Finally, S_{size} is defined as:

$$S_{\text{size}}(a,b) = \frac{|I| - |a| - |b|}{|I|}, \quad (3)$$

where I is the whole image and $|x|$ is the number of pixels in region x. With this component we encourage smaller regions to be merged first.

van de Sande *et al.* [12] has shown that it is highly beneficial to use multiple segmentations to generate a representative set of candidates for object locations. Therefore, we run the hierarchical segmentation algorithm multiple times with various parameters for initial over-segmentation. Particularly, we use the following settings for Felzenszwalb and Huttenlocher algorithm [15]: (0.8,100,100) and (0.8,200,200), where the first number is a smoothing parameter σ, the second is a threshold k, and the last is a minimum region size in pixels. Furthermore, we run [15] in four different colour spaces: RGB, HSV, Opponent Colour, and normalized RGB. van de Sande *et al.* [17] show that these colour spaces have different invariance properties, and therefore they lead to different initial over-segmentations. As a result we obtain eight different hierarchical segmentations. For further analysis we take only segments which are larger than 10% of the image width/height. In evaluation Section 4 we will show how the number of considered hierarchies influences the accuracy of salient object detection task.

3.2. Center-surround saliency

A common object characteristic is its differentiation from the background [18,1]. Furthermore, image regions which deviate from their surroundings are likely to attract attention [13,19]. Thus, ranking image segments based on their deviation from the immediate surrounding reflects the plausibility of the segment to cover a complete object and at the same time to attract attention.

With center-surround saliency, we measure the difference between segment and its surrounding by calculating a χ^2 distance of their opponent colour histograms. As surroundings, we consider pixels within an extended bounding box but outside the segment, where the extended box has an area twice larger than an area of a tight bounding box.

3.3. Integrated saliency

An object also attracts attention when it contains rare details, which we refer to as integrated saliency. We follow [13] and equate rarity of a local patch of an image to its informativeness in Shannon sense. Particularly, we measure how much information is present locally at each pixel as defined by the whole image content. To describe the image content we estimate visual word distribution and colour distribution, where the former captures image texture.

To calculate visual words, we use the fast framework of [20] with standard settings [21,12,17]. Particularly, we use the intensity-based SIFT descriptor which covers an image patch of 24×24 pixels. We do not normalize SIFT to a unit vector in order to retain contrast information. To create a visual vocabulary we quantize 250,000 randomly selected SIFT descriptors into clusters using K-means. Our vocabulary consists of 4096 visual words. To estimate the visual word distribution, we calculate frequencies of all visual words and spread them out over the patch they cover.

To estimate the colour distribution we convert an image to the opponent colour space. In the opponent colour space colour channels are uncorrelated. Therefore, following a naive Bayesian approach, we combine distributions of different channels by multiplication. As we do not use colour information in visual word calculation, the

information which is encoded in the visual word distribution is complementary to the information which is encoded in the colour distribution. Again, based on a naive Bayesian approach, we combine both distributions by multiplication.

There is strong evidence that the central part of an image attracts spatial attention [22,13]. Therefore, we add in our analysis a central bias *CB*, being a Gaussian blob centered in the middle of the image with a standard deviation σ equals to the image size. This increases the saliency of the central part of the image. Thus, for the two similar objects, the saliency of the object situated closer to the center of the image will be higher than the saliency of the object which is closer to the borders of the image.

Our final saliency map can be described as follows. If a pixel i is related to a visual word w_i, has colour $[c_{1i}c_{2i}c_{3i}]$ in the opponent colour space, and the central bias at i is $CB(i)$ then the saliency of this pixel is

$$A_i = -log(P(w_i) * P(c_{1i}) * P(c_{2i}) * P(c_{3i})) + CB(i). \quad (4)$$

To measure the integrated saliency we sum up values of the saliency map obtained by Eq. (4) within each segment. However, as all integrated saliency scores are positive, the segment which covers the whole image will always get the highest score. Therefore we threshold the saliency map and retain 50% of the most salient pixels being positive whereas the rest become negative. Note that, as we will show in evaluation Section 4, by varying the threshold we make a trade-off between precision and recall.

3.4. Selection of the most salient proto-object

We generate a set of proto-objects as described in Section 1. Then we estimate center-surround and integrated saliency of each proto-object. Center-surround saliency measures how much a segment pops out from the surrounding. Integrated saliency indicates if there are unique details within segment. These two measurements are complementary to one another. Therefore, we combine both measurements by summing up center-surround saliency and integrated saliency scores of each segment to get the final score. The segment with the maximal final score is selected as the most salient object.

3.5. Implementation details

In this paper we use the following parameter settings. To generate a set of proto-objects we run 8 different hierarchical segmentations as described in Section 3.1. This results in approximately 1200 segments per image. To calculate center-surround saliency (see Section 3.2), as a surrounding of an proto-object we consider a rectangle around the proto-object which is 2 times larger than a tight bounding box around the proto-object. To calculate integrated saliency (see Section 3.3), we threshold the saliency map from Eq. (4) and retain 50% of the most salient pixels being positive whereas the rest become negative. As a central bias we use a Gaussian blob centered in the middle of the image with a standard deviation σ equals the image size.

4. Evaluation

We test our method on the dataset from [1], where the task is to detect a bounding box around the most salient object in an image. The dataset consists of 5,000 colour images with manually labelled rectangles around the most salient object drawn by nine users. We construct the ground truth by selecting the rectangle around the salient object based on the majority agreement of all users. As our method detects the most salient segment, we report a tight bounding box around it. We follow the standard procedure [1,7,8] and calculate precision, recall, and F-measure ($\alpha = 0.5$) to evaluate the proposed method.

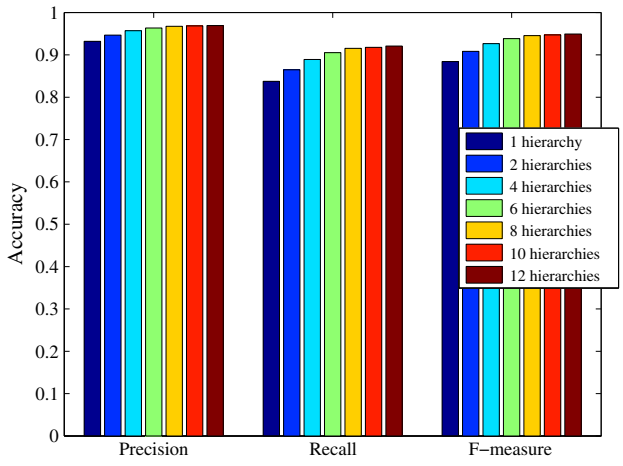

Fig. 6. The theoretical upper boundary of the proposed method: we calculate how accurately of the generated set of proto-objects covers the ground truth salient object when using an increasing number of hierarchical segmentations.

4.1. Hierarchical segmentation

The performance of the proposed method is upper-bounded by the accuracy of the hierarchical segmentation algorithm: if a generated set of proto-objects does not contain a segment outlining the ground truth salient object, our method will not be able to detect the salient object correctly. Therefore, we start with a theoretical experiment to evaluate the potential of the hierarchical segmentation algorithm. For each image we select the segment with the highest F-measure given the ground truth. Hence, we measure how well the hierarchical segmentation algorithm segments the ground truth salient objects. Moreover, to investigate how a combination of several hierarchical segmentations influences the results, we test different number of segmentations. The results are presented in Fig. 6. Clearly, using several hierarchical segmentations we can potentially achieve much better accuracy. In this theoretical setting the method reaches F-measure of 94.91% when 12 hierarchical segmentations are combined, and only F-measure of 88.42% when one hierarchical segmentation is considered. However, the accuracy is saturated once 8 or more hierarchical segmentations are used: the F-measure of 94.54% is not significantly larger than 94.91%, which corresponds to 8 and 12 hierarchies respectively. Thus, in the rest of this paper, we use 8 hierarchical segmentations to generate a set of proto-objects. As Fig. 6 shows, 8 hierarchies corresponds to high precision of 96.74%, which indicates that in most cases the algorithm succeeds to generate a segment which accurately separates a ground truth salient object from its surrounding. Furthermore, it covers a ground truth salient object adequately well, as we reach a recall of 91.53%.

Table 1
Evaluation of each component of the proposed method and comparison with the state-of-the-art saliency maps of [13,14] when applied within our framework. For the integrated saliency (Int. sal.) methods VW stands for visual words, C for colours, and CB for central bias. C-S sal. stands for the center-surround saliency.

	Without central bias			With central bias		
Method	Precision	Recall	F-measure	Precision	Recall	F-measure
Saliency [13]	77.48%	79.67%	75.44%	81.57%	82.06%	79.38%
Saliency [14]	66.69%	82.13%	68.77%	78.18%	82.11%	77.06%
Int. sal.: VW	77.07%	79.90%	75.12%	81.29%	82.25%	80.29%
Int. sal.: C	77.77%	82.19%	76.68%	83.36%	83.24%	80.91%
Int. sal.: CB	–	–	–	77.29%	79.59%	75.86%
Int. sal.: VW + C	79.80%	83.13%	78.20%	83.52%	83.91%	81.25%
C-S sal.	79.17%	60.61%	68.37%	–	–	–
Combined sal.	83.65%	82.99%	80.83%	87.61%	82.97%	83.65%

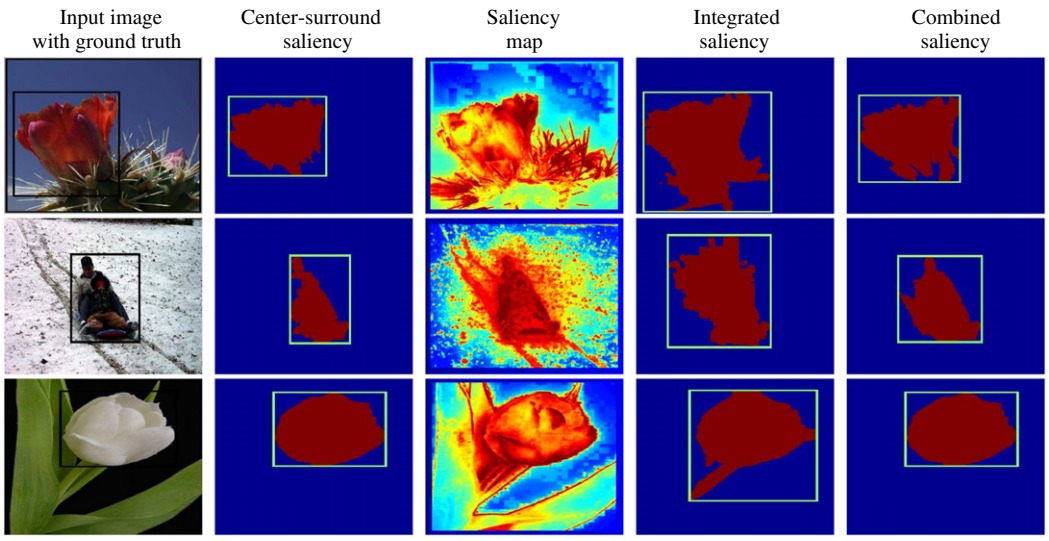

Fig. 7. Examples when center-surround saliency corrects errors made by integrated saliency.

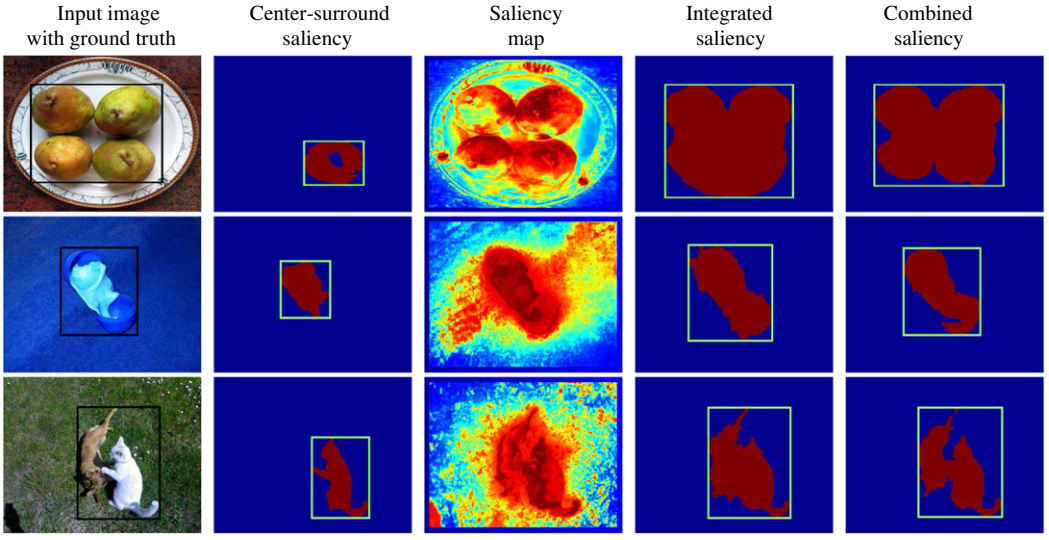

Fig. 8. Examples when integrated saliency corrects errors made by center-surround saliency.

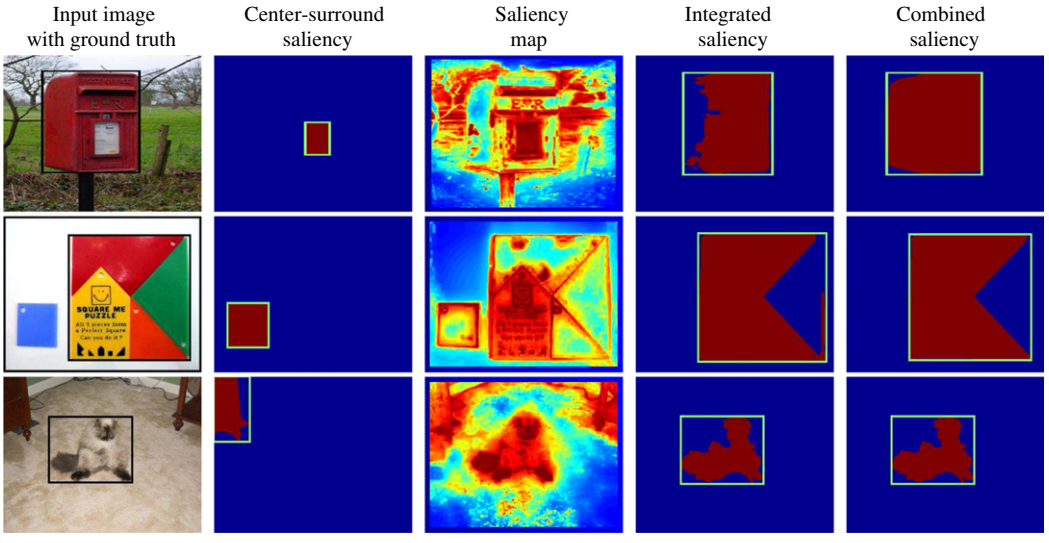

Fig. 9. Examples when integrated saliency corrects errors made by center-surround saliency.

4.2. Quantitative analysis of the results

We compare our integrated saliency to the state-of-the-art saliency maps of [13] and [14]. We compute both saliency maps of [13] and [14] using software provided by the authors. To evaluate these saliency maps on the task of salient object detection, we inserted them into our framework in the same way as described in Section 3. Practically, we took saliency maps of [13] and [14], thresholded them to leave 50% of the most salient values being positive, and calculated the sum of saliency within each segment. Furthermore, to estimate the contribution of visual words and colours to the proposed integrated saliency measurement, we evaluate each component separately. Additionally we evaluate the contribution of the central bias.

The results are shown in Table 1. As expected and has been observed in many studies [23,13,22,24], all methods perform better when combined with the central bias. The increase in precision suggests that with the central bias smaller and more accurate proto-objects are selected. Indeed, with the central bias, the emphasis shifts to selecting the salient image regions which are closer to the center of the image. This is beneficial as salient objects tend to occur more often in the middle.

Table 1 shows that each component of the integrated saliency already gains quite a good performance: visual words-based component reaches F-measure of 80.29%, and colour-based integrated saliency has F-measure of 80.91%. Moreover, our combined integrated saliency with F-measure of 81.25% outperforms both saliency maps of [13] and [14], which have F-measure of 79.38% and 77.06%, respectively. A significance of the advantage of the proposed integrated saliency is confirmed by a two-tailed t-test. We obtained a p-value of 10^{-5} for the saliency map of [13] and p-value of 10^{-9} for the saliency map of [14].

Center-surround saliency alone achieves F-measure of 68.37%, see Table 1. As described in the previous section, our center-surround saliency measures the distinctiveness of the whole segment with respect to its surrounding. This measure tends to emphasize both a segment with an object which differs from the background, as well as a segment with a background which differs from the surrounding objects, like a piece of the sky surrounded by trees. Thus, segments with distinctive background distract the center-surround saliency from the salient object, causing this lower F-measure.

When we combine center-surround and integrated saliency, the performance of our method reaches F-measure of 83.65%, see Table 1. This indicates that these two types of saliency measure complementary characteristics of saliency of an object and hence improve each other, which we will demonstrate in the next section.

4.3. Qualitative analysis of the results

Fig. 7 shows typical examples when center-surround saliency corrects errors made by integrated saliency. If the immediate surroundings of an object contain rare image details, integrated saliency tends to favour segments which contain both the object and salient parts of the surrounding. For example, it selects the flower and part of the leaves, the sleigh and part of the trail, and the tulip together with the stalk. However, such salient details typically do not belong to the object appearance. Therefore, center-surround saliency helps to find the correct borders of the salient object.

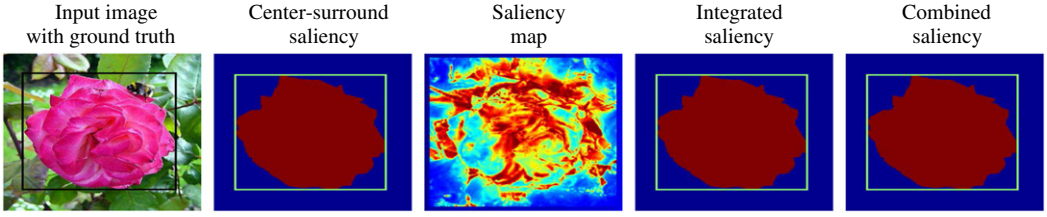

Fig. 10. Examples of mistakes.

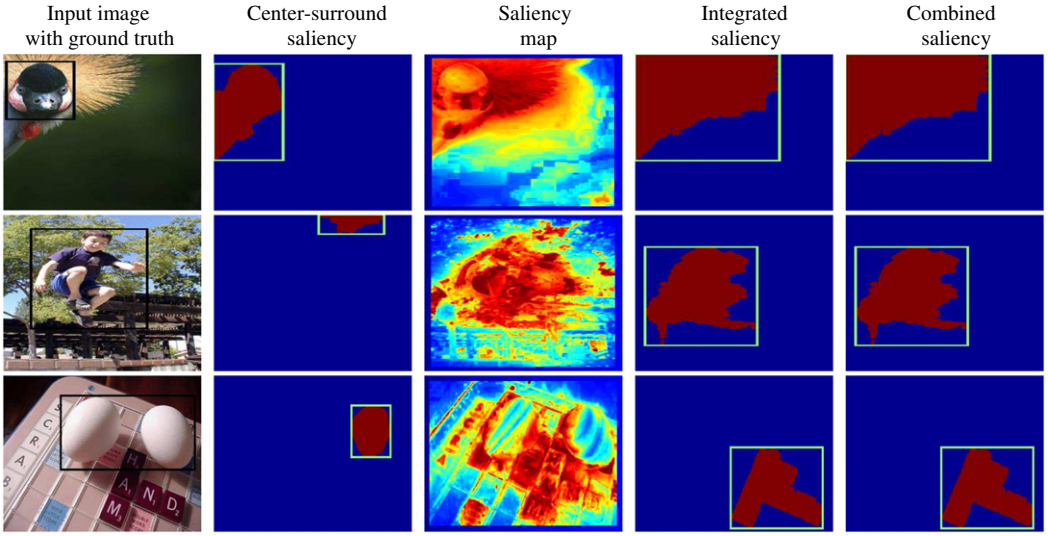

Fig. 11. Illustration of multiple salient objects detection. In the first row, 30 most salient detected proto-objects are enclosed in red frames. In the second row near-duplicate proto-objects are eliminated using non-maximum suppression.

Table 2
Comparison with the state-of-the-art salient object detectors.

Method	Precision	Recall	F-measure
Liu *et al.* [1]	83.00%	82.00%	80.00%
Valenti *et al.* [8]	84.91%	76.19%	79.19%
Marchesotti *et al.* [7]	84.50%	87.80%	85.50%*
Proposed method	87.61%	82.97%	83.65%

* Note that [7] uses a leave-one-out strategy to train their method on the dataset.

In contrast, Fig. 8 shows examples when integrated saliency corrects errors made by center-surround saliency. If there is a group of salient objects in the picture, center-surround saliency tends to choose a segment with only one salient object. For example, it selects one fruit where there are four pears, only the white milk in the picture with two bowls, and only one cat where there are two playing pets. However, such missed objects usually correspond to high energy values of the saliency map. Thus, in this case, integrated saliency helps to find the rest of the salient objects. At other times, the most distinctive from the surrounding segments are not the most salient, as it is shown in Fig. 9. The combination of both types of saliency solves this confusion.

For pictures with a single prominent object on a simple background the same segment is usually chosen by center-surround and integrated saliency, as shown in Fig. 10. Finally, Fig. 11 illustrates typical errors of our method. In the first example with an ostrich, our method selects the whole head as salient, in contrast to users, who have found salient only the muzzle. In the picture with a jumping boy integrated saliency roughly detects a boy, however, center-surround saliency does not succeed to adjust borders. In the last example, we miss the whole ground truth salient object and detect the segment with letters as salient. Here the high-level knowledge is required to avoid an error, as the eggs become salient because they are not part of the game.

4.4. Comparison with the state-of-the-art

We compare our method with state-of-the-art salient object detectors [1], [8], and [7]. As all these methods are evaluated on the same dataset [1], we directly report their results given in the original papers.

Both methods by Liu *et al.* [1] and Valenti *et al.* [8], although using different features, estimate pixel-based saliency map and then localize salient object as determined mostly by the structure of this map. In contrast, our method explicitly takes into account the information about image proto-objects while assessing object saliency. Furthermore, the central bias is incorporated in both methods. Liu *et al.* [1] explicitly add central bias in their color spatial distribution feature. Although Valenti *et al.* [8] do not have an explicit central bias, their measure does favour objects in the middle. They calculate an isocenteric saliency, where isocenters from curved regions count more heavily. This means that isocenters are more prominent when the curvature that generates them is close by. Therefore, image regions near the border and the corners have, a priori, less chance of becoming salient as there are just less pixels in the neighbourhood which can generate an isocenter response.

The results are shown in Table 2. Methods [1] and [8] have F-measure of 80.00% and 79.19%, respectively, while our method achieves F-measure of 80.91% when only the colour-based integrated saliency is used, see Table 1. Moreover, our combined saliency without central bias has F-measure of 80.83%, whereas the full method with F-measure of 83.26% significantly outperforms both [1] and [8].

The method by Marchesotti *et al.* [7] classifies each image patch as belonging to salient object or its background, where a classifier is trained on the K most similar images which have ground truth bounding boxes around salient objects. Their method outperforms our approach and achieves an F-measure of 85.50% when a leave-one-out strategy is used to train the classifier, see Table 2. By its nature, [7] requires an annotated dataset with a wide variety of content and its results strongly depend on the quality of nearest images retrieval. In contrast, our method does not require any learning.

In the task of salient object detection, it is essential to accurately locate the position of a salient object. As discussed previously in [1] and [8], a dummy system which simply chooses the whole image as a salient object will achieve a recall of 100%. This means that high precision is more important than high recall. When considering precision, our method significantly outperforms all evaluated methods by achieving the highest precision of 87.61%.

4.5. Multiple salient objects detection

The proposed framework is not limited to the detection of a single salient object. In general, any number of the most salient proto-objects can be selected. Depending on the application, the desirable number of salient objects can be defined in advance. If the number of salient objects is not known, all proto-objects with saliency above a certain threshold can be selected. In the first row of Fig. 12 bounding boxes around 30 most salient proto-objects are shown for several images. In the second row of Fig. 12 we demonstrate how extraction of the same object several times can be avoided. Particularly, we use non-maximum suppression to eliminate near-duplicate proto-objects with a higher overlap than 50% with proto-objects that are more salient.

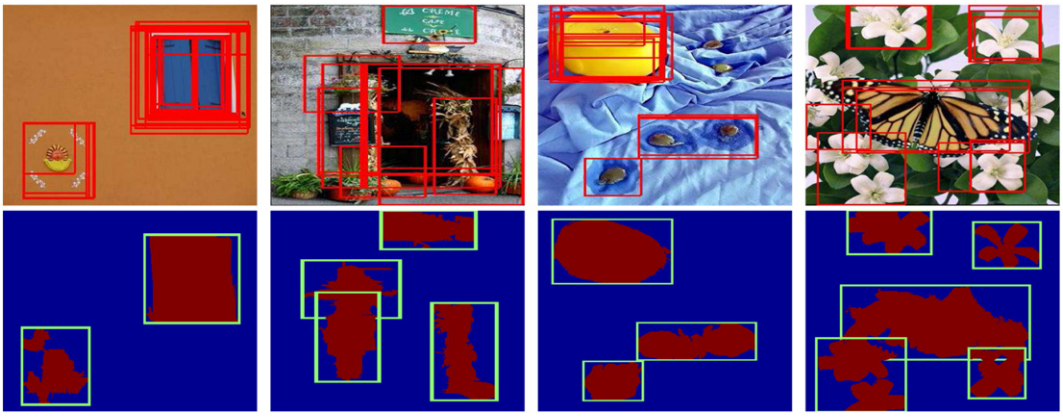

Fig. 12. Influence of the number of hierarchical segmentations to the accuracy.

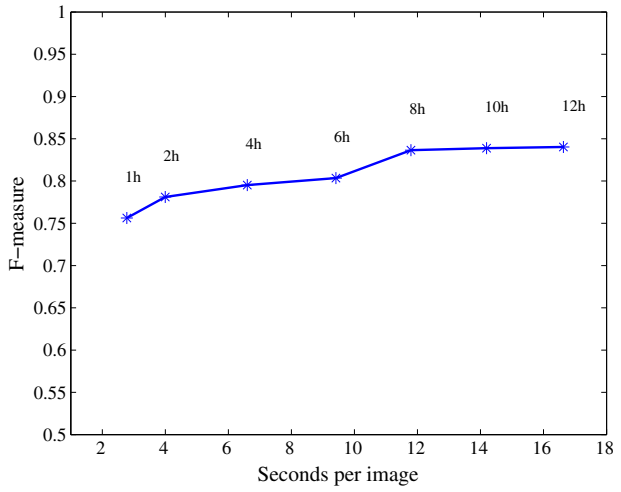

Fig. 13. Influence of the size of the surrounding and the value of the threshold to the accuracy of the proposed method.

4.6. Parameter evaluation

Here we investigate the influence of the settings of parameters of the proposed method to the accuracy and computational efficiency. Our method can be divided into two parts: (1) extraction of proto-object, and (2) estimation of saliency of proto-objects. In the proto-object extraction part, there is only one parameter which controls the number of hierarchical segmentations used in the method. The more hierarchical segmentations are used, the more proto-objects are generated. Therefore, as it has been shown in the theoretical experiment in Section 1, there is a better chance to detect the ground truth salient object. From another side, the more proto-objects are considered, the longer is the computational time. Thus, the number of hierarchical segmentations is a trade-off parameter between accuracy and computational efficiency of the method. Fig. 13 illustrates how both the computational time, which is required to process one image, and F-measure are growing as a function of the number of hierarchical segmentations. Experiments were performed on an Intel(R) Xeon(R) CPU E5620 @ 2.40 GHz. When the speed is critical, we recommend to use 4 hierarchical segmentations which corresponds to F-measure of 79.52% and requires only 6.5 s per image. Otherwise, it is optimal to use 8 hierarchical segmentations which corresponds to F-measure of 83.65% and requires 11.8 s per image.

In the estimation of saliency of proto-objects part, there are 3 parameters: the size of surroundings, the threshold of saliency map, and the size of center bias. Theoretically, the first parameter is largely orthogonal to the rest: the size of the surroundings is used for center-surround saliency and determines if the segment should be compared with only its immediate surroundings or with the larger area up to the whole image. We do expect this to have little influence on results. The threshold parameter, together with the size of central bias, is part of the integrated saliency. The threshold controls which portion of the saliency map retains positive values, whereas the rest of the saliency map is converted to negative values. Hence the threshold influences the size of the most salient segment. We expect this parameter to have a strong influence on the accuracy. The size of center bias controls the emphasis of the central part of the image. The stronger is the central bias, the higher is the saliency of the central part of

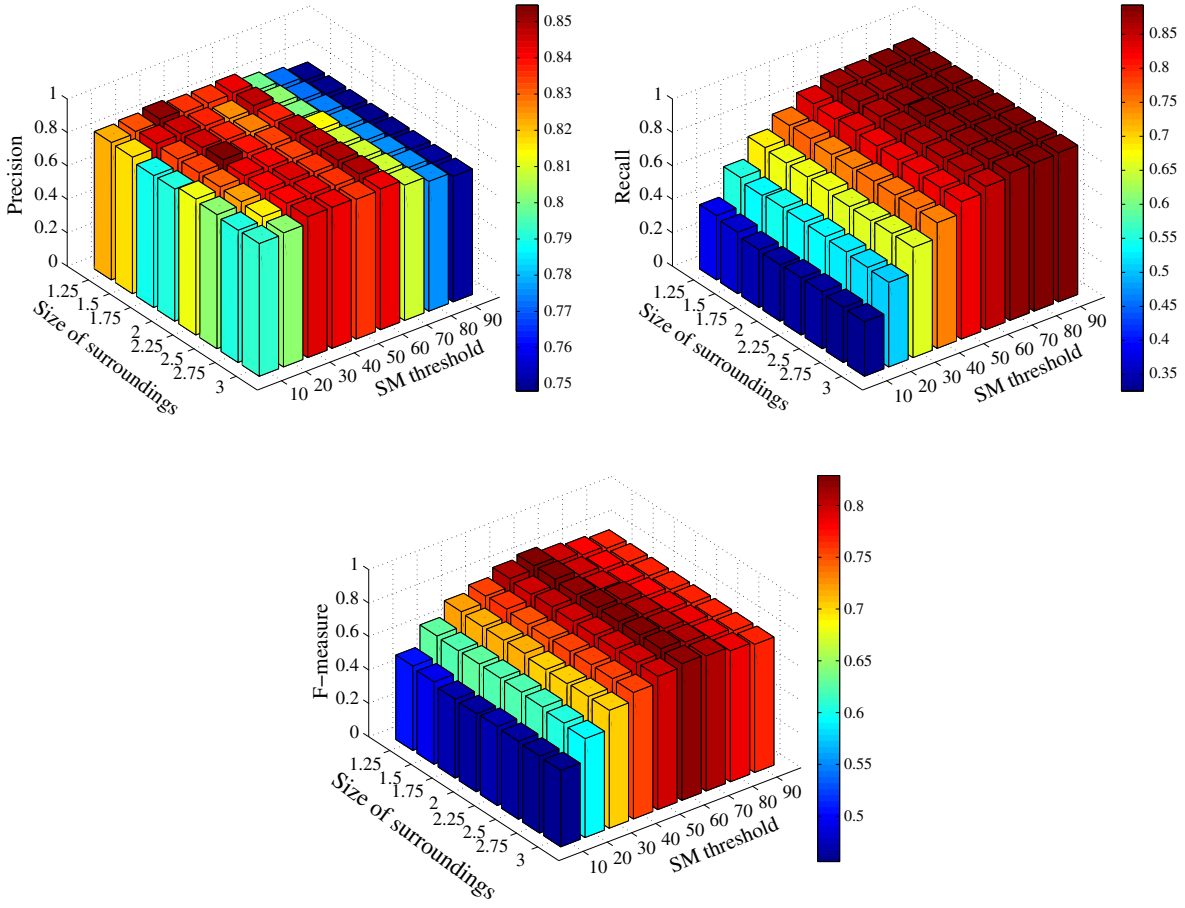

Fig. 14. Influence of the size of the central bias and the value of the threshold to the accuracy of the proposed method.

the image in comparison with the rest of the image. Many papers report that salient objects tend to appear near to the center of the image [23,24]; thus, as we have shown in Table 1, the central bias usually boosts the performance.

Actually, there are more parameters of the method. We use the bag-of-words paradigm for image representation when calculating the integrated saliency. This paradigm includes parameters such as the size of the SIFT descriptor, the rate in which SIFT descriptors are sampled, and the size of the visual vocabulary. However, we take this approach from the shelf and use it as it is with all standard settings. Therefore, we do not evaluate these parameters.

We start with testing how different settings of the size of surroundings and the threshold of saliency map jointly influence the accuracy of the proposed method. Results presented in Fig. 14 confirm independence of these two parameters. The size of the surroundings hardly influences the accuracy, although smaller surroundings (1.25 and 1.5 of a bounding box size) are preferred. In contrast, the threshold of saliency map has a large impact on the performance. Therefore, we recommend to optimize this parameter, while any choice within a range from 50% to 70% is reasonable. Importantly, it can be observed that the maximum of the threshold is 60% regardless of the size of surroundings, which shows that this parameter can be tuned independently of any choice of surroundings. In our experiments, the threshold of 60% corresponds to the highest F-measure of 84.07%, whereas F-measures corresponding to 50% and 70% have very close values of 83.65% and 83.86% respectively.

Furthermore, we analyze the joint influence of the threshold of saliency map and the size of center bias to the accuracy. Fig. 15 demonstrates that these two parameters depend on each other. The range of the threshold values from 40% to 70% corresponds to the highest F-measure. Note, that this set of thresholds contains the range of 50–70%, which we have found optimal in the previous paragraph (see Fig. 14). The central bias with σ equals to 75% of image size brings the best performance. However, within the optimal range of threshold values, the influence of the central bias remains stable when σ varies from 75% to 125% with small drops in performance towards larger values.

To conclude, we recommend the following settings: (1) use 8 hierarchical segmentations to generate a rich variety of proto-objects and reach high accuracy, (2) take into account rather small surroundings of a proto-object to calculate center-surround saliency, (3) preferably jointly optimize the threshold value of the saliency map together with the size of central bias (if not applicable, set the threshold within 50–70% range and consider central bias with σ equals 75–125% of image size)."

5. Conclusions

We have proposed a novel framework for the task of salient object detection inspired by the object-based visual attention theory. We assume a proto-object being a unit of attention and argue that notion of an object should be taken into account while assessing object saliency. Furthermore, we consider two types of object saliency: center-surround saliency measures how an object differs from its surrounding, and integrated saliency measures how many rare details are within an object. We demonstrate that both types of saliency have complementary characteristics, and the combination improves the performance. The proposed method achieves state-of-the-art results on a well-known benchmark.

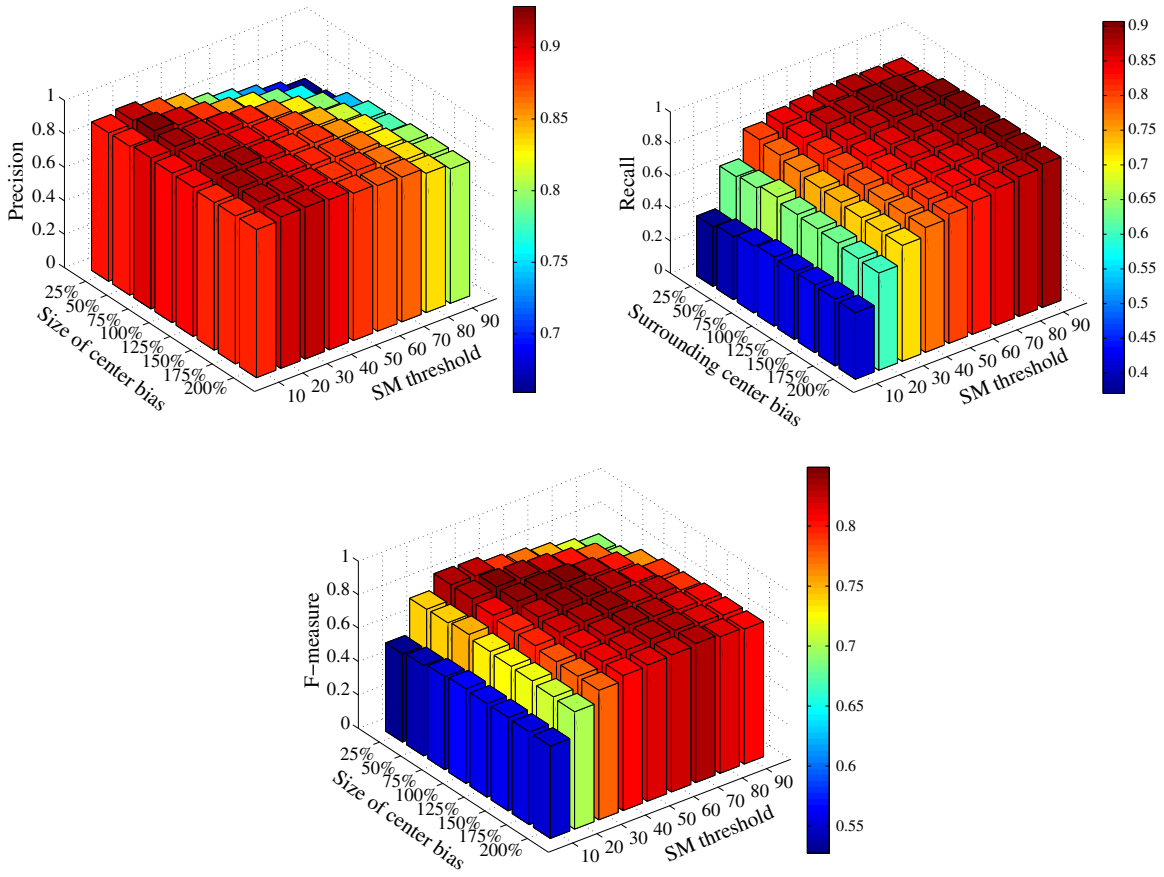

Fig. 15. Influence of the size of the central bias and the value of the threshold to the accuracy of the proposed method.

References

[1] T. Liu, J. Sun, N.N. Zheng, X. Tang, H.Y. Shum, Learning to detect a salient object, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.
[2] C.W. Eriksen, J.D. St James, Visual attention within and around the field of focal attention: a zoom lens model, Percept. Psychophys. 40 (4) (1986) 225–240.
[3] A.M. Treisman, G. Gelade, A feature-integration theory of attention, Cogn. Psychol. 12 (1) (1980) 97–136.
[4] B.J. Scholl, Objects and attention: the state of the art, Cognition 80 (1–2) (2001) 1–46.
[5] R. Rensink, Seeing, sensing, and scrutinizing, Vision Res. 40 (10–12) (2000) 1469–1487.
[6] D. Walther, C. Koch, Modeling attention to salient proto-objects, Neural Netw. 19 (9) (2006) 1395–1407.
[7] L. Marchesotti, C. Cifarelli, G. Csurka, A framework for visual saliency detection with applications to image thumbnailing, in: International Conference on Computer Vision, 2009, pp. 2232–2239.
[8] R. Valenti, N. Sebe, T. Gevers, Image saliency by isocentric curvedness and color, in: International Conference on Computer Vision, 2009, pp. 2185–2192.
[9] C.H. Lampert, M.B. Blaschko, T. Hofmann, Beyond sliding windows: object localization by efficient subwindow search, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[10] C. Rother, V. Kolmogorov, A. Blake, Grabcut: interactive foreground extraction using iterated graph cuts, in: International Conference on Computer Graphics and Interactive Techniques, 2004, p. 314.
[11] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (11) (1998) 1254–1259.
[12] K.E.A. van de Sande, J.R.R. Uijlings, T. Gevers, A.W.M. Smeulders, Segmentation as selective search for object recognition, in: International Conference on Computer Vision, 2011, pp. 1879–1886.
[13] N. Bruce, J. Tsotsos, Saliency, attention, and visual search: an information theoretic approach, J. Vis. 9 (3) (2009) 1–24.
[14] X. Hou, L. Zhang, Saliency detection: a spectral residual approach, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.
[15] P.F. Felzenszwalb, D.P. Huttenlocher, Efficient graph-based image segmentation, Int. J. Comput. Vision 59 (2) (2004) 167–181.
[16] J.M. Geusebroek, R. van den Boomgaard, A.W.M. Smeulders, H. Geerts, Color invariance, IEEE Trans. Pattern Anal. Mach. Intell. 23 (12) (2001) 1338–1350.
[17] K.E.A. van de Sande, T. Gevers, C.G.M. Snoek, Evaluating color descriptors for object and scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1582–1596.
[18] B. Alexe, T. Deselaers, V. Ferrari, What is an object, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 73–80.
[19] D. Gao, V. Mahadevan, N. Vasconcelos, On the plausibility of the discriminant center-surround hypothesis for visual saliency, J. Vis. 8 (7) (2008) 1–18.
[20] J.R.R. Uijlings, A.W.M. Smeulders, R.J.H. Scha, Real-time bag of words, approximately, in: International Conference on Image and Video Retrieval, 2009, pp. 1–8.
[21] M. Marszałek, C. Schmid, H. Harzallah, J. van de Weijer, Learning object representations for visual object class recognition, in: Visual Recognition Challenge workshop, in conjunction with ICCV, 2007, pp. 681–688.
[22] B.W. Tatler, The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions, J. Vis. 7 (14) (2007) 4.
[23] T. Judd, K. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look, in: International Conference on Computer Vision, 2009, pp. 2106–2113.
[24] M. Spain, P. Perona, Measuring and predicting object importance, Int. J. Comput. Vision 91 (1) (2010) 59–76.