A Closer Look at Context: From Coxels to the Contextual Emergence of Object Saliency

Rotem Mairon and Ohad Ben-Shahar

Dept. of Computer Science, Ben-Gurion University of the Negev, Beer Sheva, Israel
{rotemra,ben-shahar}@cs.bgu.ac.il

Abstract. Visual context is used in different forms for saliency computation. While its use in saliency models for fixations prediction is often reasoned, this is less so the case for approaches that aim to compute saliency at the *object* level. We argue that the types of context employed by these methods lack clear justification and may in fact interfere with the purpose of capturing the saliency of whole visual objects. In this paper we discuss the constraints that different types of context impose and suggest a new interpretation of visual context that allows the emergence of saliency for more complex, abstract, or multiple visual objects. Despite shying away from an explicit attempt to capture "objectness" (e.g., via segmentation), our results are qualitatively superior and quantitatively better than the state-of-the-art.

1 Introduction

The remarkable ability of the visual system to rapidly attend towards salient stimuli enables humans to effortlessly filter visual input and allocate attentional resources differentially to salient regions. The computational prediction of this outcome can facilitate numerous applications in both the analysis of images (i.e., in computer vision) and their synthesis (i.e., in graphics). For example, the need to adjust visual context to a range of display devices has motivated image/video retargeting and content-aware resizing techniques that rely on saliency prediction [12,49,4,34,19]. A capacity to predict what is salient or not has also spared much computational resources in image classification [39], retrieval [13], object recognition [43] image and video compression [15,50], and served various other applications such as image thumbnailing [34,45], visualization and symmetrization [47,18,42] and object segmentation [21,30].

Judging by this variety of applications, the abundance of existing work on saliency computation and the need for perceptually-consistent and accurate saliency predictions are not surprising. We begin this work by taking a closer look at the mechanisms used to compute saliency and to examine the constraints and limitations they may pose on the computational process. Central to our exploration is the concept of "context" and part of our goal is to argue that it (i.e., context) alone is a sufficient substrate from which saliency can fully emerge. As we show later, despite using this single building block, our saliency results exceed state-of-the-art performance from methods that employ diverse set of additional tools and mechanisms.

D. Fleet et al. (Eds.): ECCV 2014, Part V, LNCS 8693, pp. 708–724, 2014.

1.1 Saliency and Context

From an ecological perspective, the saliency of a constituent in a visual scene is the degree to which it demands the allocation of computational (attentional) resources in order to better inquire its role in the visual stimulus. In practice, as is also acknowledged in both perceptual [46,38,16] and computational [29,14] accounts, saliency is strongly influenced (and often fully determined) by the degree to which the constituent stands out from its context. Combining the two, the saliency of a visual constituent cannot be determined without knowledge or understanding of the context in which it is embedded. Interestingly, this constituent-context duality has taken different forms in previous research of saliency computation.

Saliency is primarily driven in a bottom-up manner, depending on low level visual cues in the visual scene. In one of the first biologically plausible computational models for controlling visual attention, Koch and Ullman [31] followed Treisman and Gelade [46] and introduced the idea of a saliency map. Visual input is first decomposed into several maps encoding early visual features. Spatial competition in terms of hierarchical center-surround differences then determines their convergence to a unique map encoding saliency at each location. Most subsequent bottom-up saliency algorithms followed this model and compute the saliency of pixel constituents based on their local context (i.e., neighborhood) at multiple scales [27,22,10,25]. Alternatively, context was also considered globally, e.g., as a smoothed version of the amplitude [23] or the phase [20] spectrum of the image. Deviations from the original non-smoothed spectrum with respect to this global context are then considered as salient locations when transformed back to the spatial domain.

In addition to its categorization as local or global, bottom-up saliency may also be viewed at the level at which it operates. Unlike the models mentioned above, that mainly act spatially in order to reproduce human visual search strategies or predict visual fixations, other methods aim at detecting saliency at the higher level of *objects*. While the (local) visual context used by the first class of methods is reasonably intuitive, the forms of visual context employed by the latter (object-level) approaches typically remain unexplained. We argue below that this somewhat obscure relationship often constrains the nature of visual objects they may capture in order to measure their saliency.

Considering the scope of saliency as discussed above, we define visual context of a constituent as follows:

Definition 1. *The visual context of a constituent is the set of visual units in the image that are used in the computational process that measures its saliency.*

This somewhat general definition intentionally lacks a particular spatial relationship between the constituent and it context. It is used in Sec. 2 to discuss the contribution of different types of visual context to detecting saliency at the object level and to point at the constraints that these types of context may impose. Then, in Sec. 3, we suggest a novel approach to visual context, which is intuitively justified and can capture object saliency for both simple, complex, and abstract objects (Fig. 1) all without explicit reference to "objectness" or the use of segmentation.

Before beginning our closer look at visual context, one disclaimer is advised. Like many others, in this work we too discuss the notion of visual context that is associated

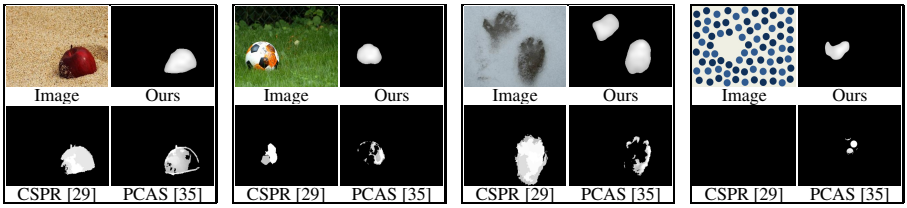


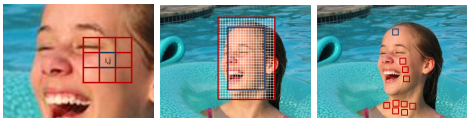
Fig. 1. Salient objects in visual stimuli can have different flavors. As is typical in virtually all benchmark databases, salient objects can be uniform singletons (panel a). However, salient objects can be multi-part and heterogeneous (panel b), they can have some multiplicity (panel c), or they can even be completely abstract (like the "hole" in panel d). By their implied notion of visual context, most computational saliency models impose certain constraints on the types of objects they can handle, with practical success limited to the simpler cases. Here we show computed saliency map (thresholded at 80%) from two state-of-the-art algorithms (CSPR [29] and PCAS [35]) and our own method. By modeling context instead of the objects we significantly reduce the constraints on the nature of objects that may be detected as salient, as is illustrated by the better assignment of saliency in all these cases.

with bottom-up saliency. But the latter may be strongly modulated or even overridden by top-down factors as well, including the experience (or expertise) of an observer or his biases due to task definition [26]. Such factors give rise to other forms of visual context and modulation of bottom-up saliency by semantic interrelations between visual objects [7,5] or the global structuring of a scene [6,41,37,40]. These types of context remain outside the scope of our present work.

2 Background and Related Work

Approaches to salient object detection embrace the same notion of a saliency map discussed above (sometimes with additional steps like segmentation) but employ different types of visual context (in the sense of the Def. 1) to compute such maps (see Fig. 2). To address the specific contribution of the types of context used we roughly categorize the different approaches into the following two groups:

Contrast-Based Saliency: In the first group are approaches that associate saliency with high contrast between local or regional structures. To measure this contrast, the computational mechanisms employ various center-surround structures. The visual constituent for which a measure of saliency is computed is regarded as the center and is spatially surrounded by its context. Some approaches define the surround component independent of visual content, e.g., as the local neighborhood of a pixel [24,48,1,32] or larger regular blocks [33]. In other approaches, the surrounding context depends on a grouping process which typically results in a superpixel representation of the image [29,11]. Apart from reducing computational costs, superpixels are preferable due to their capacity to preserve locally coherent structures



(a) (b) (c)

Fig. 2. Different types of visual context (marked in red) of a visual constituent (marked in blue). **(a)** The local neighborhood of a pixel. **(b)** Pixels at the surround of a larger scale region. **(c)** *k*-nearest neighbors of a patch.

(unlike pixels or predefined blocks). To a certain extent, these structures facilitate meaningful central constituents when measuring contrast and therefore are more suitable for saliency assignment.

Rarity-Based Saliency: The second group of approaches consider saliency as distinctness or rarity. Intuitively, these may signal the importance of a visual constituent compared with the redundancy of recurring visual information. Often in this approach the context is a global representation of the entire visual input. A constituent is then considered salient if its representation does not conform with the context. For example, such a representation may be the image mean color vector that is used as reference to measure the saliency at all other pixels [2,4]. Alternative representation has considered a smoothed version of the phase spectrum [28] in order to suppress non-salient components in the original spectrum and thus highlight salient locations after transforming back to the spatial domain. In a somewhat related way, image patches that are highly dissimilar to their *k*-nearest neighbors were considered salient as this indicates their dissimilarity to all other patches [19,11]. Recently, this measure of dissimilarity has been shown oblivious to patch statistics, leading to a new measure based on the distance of each patch to the average patch along the principal components of the patch distribution [35].

An important factor in approaches from both of the groups above is the scale at which saliency is computed. When the context is predefined as the surround in a certain center-surround structure or as a global description of the visual input, its scale may be selected arbitrarily. In case it is determined by a grouping process, the scale may be influenced by different input parameters. However, in both cases there is no single appropriate scale. Tightly localized context would essentially capture edge information while context of excessive spatial scale may falsely signal non-salient areas and incorporate visual information whose relevance to the saliency of a visual constituent is unclear. Thus, the saliency map is often a combined result of computations across multiple scales.

Other complexities that visual objects may exhibit pose additional constraints to the nature of visual objects that may be captured during saliency computation. Indeed, the implicit motivation underlying contrast-based saliency is the possibility that at a certain scale the center part of the center-surround structure will capture the object to allow

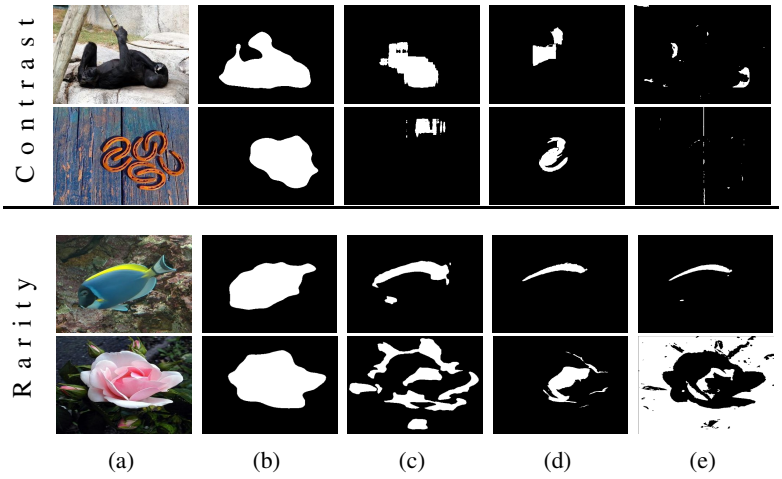


Fig. 3. Binarized saliency maps demonstrate the challenges in capturing whole salient objects by contrast (top) and rarity (bottom) based approaches. The two leftmost columns in each category show example images and our maps. **Constrast:** Saliency maps in columns c and d are generated as part of saliency computation algorithms, but are not their final output (which includes additional steps). They are shown here to demonstrate how capturing large or discontiguous objects is constrained when relying on regional center-surround. In column c computation is based on rectangular structures of varying size and aspect ratio [32] whereas in column d neighboring superpixels were used to estimate contrast [29]. The constraints are even more restrictive when only local considerations are involved [1] as shown in column e. **Rarity:** The challenge remains when relying on rarity aspects of saliency, as demonstrated by the maps in columns c-e [19,35,14]. When the object consists of multiple parts, only those with rare appearance are detected. The bottom map in panel e demonstrates how a large object may render the appearance of its surrounding more rare and therefore more computationally salient.

the comparison of its appearance against its surroundings. This implies that the object is expected to be compact and spatially continuous. Compactness and spatial continuity may not be required for rarity-based saliency, which assumes that the target object constitutes few units with rare visual properties with respect to the entire visual input. However, this approach ignores spatial relations between elements forming the context and may not account for figure-ground relations. In fact, when relying on rarity, the surrounding of a visual object may be considered more salient when the object is larger. The rarity aspect of saliency is also challenged when it comes to considering composite/heterogeneous objects. In these cases, different parts of a salient object may be assigned very different saliency values (see Fig. 3).

The limitations just discussed have led many scientists to use additional information and computational processes to possibly capture the nature of visual objects. Often, saliency maps are used as input to subsequent segmentation processes such as adaptive thresholding [2] fuzzy-growing [33], compactness and density analysis [24], and

iterative region expansion [52]. Additional considerations are configural cues such as convexity [48] or closure [29], or higher-level factors such as objectness [11] and visual organization priors [35,19]. In other cases, the additional information used is more explicit and extracted directly from a collection of images (e.g., [32]).

While many of the approaches above indeed improve the original saliency mapping, The difficulty of modeling the nature of visual objects often leads to ad hoc methods that blur the distinction between bottom-up saliency and its applications in subsequent computations. In this work we propose a completely different approach. Instead of trying to capture the object, we put the emphasis on modeling the context that leads to visual saliency. As we show later, this paradigm shift leads to superior saliency results even if no additional object-specific information or computational processes (like segmentation) are employed.

3 Modeling Visual Context to Compute Saliency

Essentially, the same fundamental question is at the basis of most approaches to saliency computation: "To what extent does a visual constituent stand out from its context". This question implies that a certain constituent is at hand when its saliency is measured or estimated. When the desired constituent is an object, this idea raises the issues described above that limit the performance. Instead of trying to capture the object, we wish to consider a somewhat dual question: "What are the characteristics of visual context which allow to consider the visual information it embeds (be it an object or not) as salient".

To answer this question, we suggest to model visual context based on the several characteristics of visual information. Given a particular representation of the units that compose it (pixels, superpixels, patches, etc...), we consider a single *context element*, or *coxel*, to be a region or a subset of the image with the following properties (see Fig. 4):

Smoothness: Nearby units that compose the coxel are expected to have similar visual appearance. The more distant the units, more leeway is allowed in their similarity.

Apathy to contiguity: A coxel may be either contiguous or not, i.e., it may constitute several distinct connected components in the image plane.

Enclosure: To qualify as a saliency coxel, the spatial layout of the context element should "enclose" (strictly or approximately) some visual information.

While many ways can be used to define elementary image units from which coxels are composed, we elect to do so via the approximately regular, boundary adhesive patches such as those obtained from the SLIC superpixels algorithm [3]. Let $V = \{v_1, \ldots, v_n\}$ be the set of all these patches. Each patch is associated with a single coxel, the latter being a subset of V with the properties outlined above. Let C be the mapping from each patch to its coxel, such that $C(v_i)$ is the coxel of patch v_i. We denote the set of all coxels by $\mathbb{C}$. Initially, $\forall i, C(v_i) = \{v_i\}$ and $|\mathbb{C}| = N$.

Let $G = (V, E)$ be the weighted *complete* graph on V, where the weight $w(E_{ij})$ of each E_{ij} reflects the *contextual gap* between its corresponding patches v_i and v_j. Two general factors affect the contextual gap – similarity in appearance and image distance. The contextual gap as a whole, and the similarity distance in particular, can be evaluated in various ways. Here we choose to use a particularly simple form that takes only the

raw color as a measure of appearance and the following blend of color and distance to express contextual gap

$$w(E_{ij}) = 1 - \left(\frac{1 - \alpha * s_{ij}}{1 + \beta * c_{ij}} \right) \tag{1}$$

where c_{ij} and s_{ij} are the appearance (color) distance and the spatial distance between the pair of patches, respectively, and α and β control their significance ($\alpha = 0.5$ and $\beta = 7$ were used). This results with contextual gaps in the range [0, 1] that are lower for edges linking similar and nearby patches and higher otherwise. The choice to express appearance similarity very simply via color only is intentional since it implies that the strength of our approach must emerge from the proposed concept of context and the derived estimation of saliency. Indeed, as we'll show, while our algorithm can accept arbitrarily sophisticated appearance measures, even the naïve one employed here already results in better than state-of-the-art saliency performance (even without endowing it with segmentation or other additional computational processes).

With the initial coxels set and pairwise contextual gaps between patches determined, our algorithm proceeds by repeatedly altering between two computational phases. The first phase enables coxels to extend by gradually merging together coxels of increasing contextual gap. The second phase accumulates saliency votes for visual information that is embedded in (i.e., enclosed by) coxels. Upon convergence, the entire image becomes a single coxel and the saliency map is finalized.

More formally, given the graph G and a predefined desired quantization level of contextual gaps $0 = w_1 < w_2 < \ldots < w_m = 1$, the steps described in Algorithm 1 (and illustrated in Fig. 5) are repeated until a single coxel is reached. In the first phase, coxels are extended by merging existing coxels by progressively relaxing the contextual gap allowed. Leveraging the smoothness property, initially only nearby and highly similar components are considered for merging. Apathy to contiguity is supported by the fact

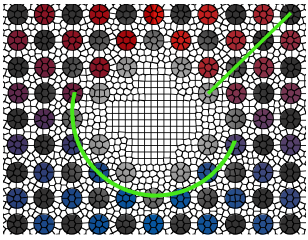


Fig. 4. The complexity and diversity of visual context that our model allows is demonstrated by this synthetic image. White, colored, and grayscale patches (superpixels) compose a scene of circles surrounding an "empty" salient region (cf. Fig. 1d). Context elements can be regarded at the level of these patches or at a higher level depicting circles and white background. Although the appearance of context units varies around the empty salient region (e.g., along the curved green path) and away from it (straight green line), at some level they should be considered as part of the *same* context element. In our approach to context this is possible due to the smoothness property and the lack of contiguity which allow context elements from different sides of the salient region to merge.

that the increased contextual gaps w_l gradually permit the merging of more distant and less similar coxels even if they are disconnected. Thus, a pair of patches v_i and v_j may (and at some point, will surely) belong to the same coxel, such that $C(v_i) = C(v_j)$.

Algorithm 1. Contextual Emergence of Saliency

$S(E_{ij}) := 0 \;\; \forall i, j = 1..n$ {Initial votes for saliency bridges}
$l := 0$
while $|\mathbb{C}| > 1$ **do**
  {Phase I: Extend coxels}
  for all E_{ij}, s.t $w(E_{ij}) \leq w_l$ and $C(v_i) \neq C(v_j)$ **do**
    $\mathbb{C} = \mathbb{C} - C(v_j)$
    $C(v_i) = C(v_i) \cup C(v_j)$
  end for
  {Phase II: Accumulate saliency votes}
  for all E_{ij} s.t $C(v_i) = C(v_j)$ **do**
    $T := \{v_k : E_{ij}$ traverses $v_k\} - \{v_i, v_j\}$
    if $|T| = |T - C(v_i)|$ **then**
      $S(E_{i,j}) = S(E_{i,j}) + 1$
    end if
  end for
  $l := l + 1$.
end while

During the second phase of each iteration, coxels that emerged up to this point are used to add saliency for the visual information they enclose. This is done by considering "visibility edges" or "saliency bridges", i.e., edges between patches of the *same* coxel that do *not* traverse another patch from that coxel. More abstractly, saliency bridges reflect interference in their associated context element and therefore suggest that visual information they traverse deserve a quota of saliency (all in the spirit of seeking the "extent to which a visual constituent stands out from its context"). The longer (i.e., more iterations) the relationship between a coxel and its enclosed region endures, the more "votes" saliency bridges will accumulate to indicate so.

It is easy to see that the algorithm always terminates. Since merging coxels reduces their total number, and since for every edge E_{ij} there exist some threshold w_l that exceeds its contextual gap $w(E_{ij})$, the iteration must end. Indeed, when $w_l = 1$ all remaining coxels merge into one final element, no saliency bridges are possible any longer, and the iteration terminates. In practice we represent saliency bridges by the image pixels they traverse and votes are accumulated in those pixels. Although one could employ different ways to obtain a dense map from the spatially distributed votes assigned to pixels, we apply a kernel density estimation [9,44] to produce the final saliency map.

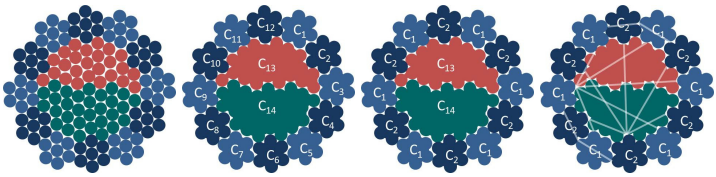


Fig. 5. Schematic depiction of the two phases of Algorithm 1. **(a)** Initial coxels (SLIC superpixels [3]) with their color-coded appearance content. **(b)** Coxels with small contextual gaps (initially, those which are very proximate and similar) are merged to larger, uniquely labeled components. Note that at this time no saliency bridges occur as any edge between two patches from the same component traverses another patch from that component. **(c)** At a future merging step, the threshold on contextual gaps is large enough to allow distant coxels to merge (implied by similar labels). **(d)** At this point, saliency bridges cross image patches from other coxels, leading to accumulation of their saliency measure. To avoid clutter, only selected number of saliency bridges are shown.

To conclude, we consider context as relevant to the saliency of a visual constituent when it exhibits certain properties that allow it to form coherently while spatially enclosing the constituent. By considering any visual information that is not part of a context element as salient, we successfully disregard issues of shape, size, contiguity, or topology, thus significantly reducing the constraints on the nature of objects that may be detected as salient (see Figs. 1 and 3). We note that the saliency bridges mechanism implicitly encourages enclosure, the third property we defined as desired. Indeed, saliency is voted for along saliency bridges, and the latter are more frequent for coxels that better enclose an image region. In addition, since saliency bridges are more likely to occur closer to the image center, an implicit centeral bias is predicted. This may in fact support the biological plausibility of the model and perhaps partially explain why humans have central bias. Finally, since coxels are apathetic to contiguity, the entire approach can capture abstract salient objects in the form of "holes" or "gaps" in a group of scattered similar elements (cf. Fig. 1).

4 Evaluation

To evaluate our model [1], we use the five datasets employed in the proposed benchmark by Borji et al. [8] and an additional dataset that was published recently by Yan et al. [51], all of which are described below.

MSRA: 5000 images of resolution 400×300. For each image, nine users annotated what they considered the most salient object by a single bounding-box.

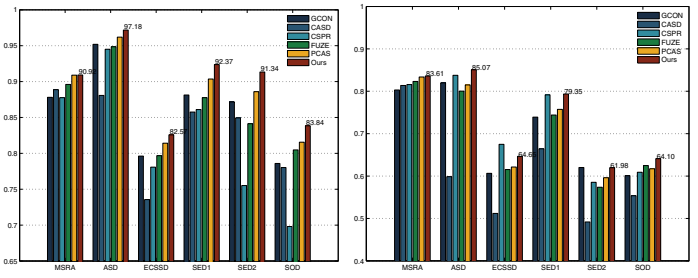
ASD: 1000 images (taken from the MSRA dasaset). For each image, a single annotator manually labeled the boundaries of a single salient object (or several of them in a few cases).

[1] Implementation will be made publicly available at `http://www.cs.bgu.ac.il/~icvl`

SED1,SED2: Each contains 100 images, of resolution $\sim 300 \times 225$. The datasets were designed to avoid ambiguities by only including images that clearly depict a single (SED1) and exactly two salient objects (SED2). Each of three annotators manually labeled the boundaries of a single or two salient objects, respectively.

SOD: 300 images of resolution 481×321, selected from the Berkeley Segmentation Dataset (BSD) [36] and labeled by seven annotators. Each annotator was shown a random subset of possible segmentations depicted as boundaries overlapped on the image and chose the segments composing salient objects by clicking on them.

ECSSD: 1000 images of resolution $\sim 400 \times 300$, taken from BSD, the VOC dataset [17] and the internet. Salient objects were manualy segmented by five annotators. However, the produced ground truth maps are binary.



(a) (b)

Fig. 6. Detection accuracy: (a) AUC scores of the "Top-4" algorithms, GCON, CASD, CSPR and FUZE, are compared with the rarity based approach recently suggested by Margolin et al. (PCAS) and our approach. On the MSRA dataset, our approach is comparable to PCAS which outperforms the "Top-4" algorithms. More significant improvements are obtained for the other four datasets. The most significant improvement is for the SED2 dataset, specifically designed to include two salient objects in every image. (b) F-Measure scores of the "Top-4" algorithms, PCAS and our approach, based on the precision-recall curve. Excluding the ECSSD dataset on which the CSPR algorithm that employs shape prior shows better scores, our approach is better than or comparable to other algorithms on all other datasets despite using nothing else but raw contextual consideration.

According to the recent benchmark by Borji et al. [8], the 4 highest scoring algorithms (henceforth, the "Top-4") to-date are FUZE [11], CSPR [29], CASD [19] and GCON [14]. Recently, Margolin et al. [35] have shown their approach (henceforth PCAS) outperforms these methods on all datasets used for the benchmark in terms of area under the ROC curve (AUC) scores. We compare our results to these five state-of-the-art algorithms, based on the same ranking used in the Borji et al. benchmark [8], both in terms of AUC scores and in terms of F-measure. Figure 6a shows AUC scores for each dataset, based on true positive rate and false positive rate, by varying a threshold from 0 to 1 on the normalized saliency maps. Our approach is comparable to PCAS on the MSRA dataset and outperforms all five algorithms on all other datasets. Interestingly, the most significant improvement is achieved on the SED2 dataset, which

includes two salient objects in every image and departs the most from the typical scenarios of single salient object around the center of the image.

Figure 6b shows the evaluation results according to the *precision-recall* curve (PR), obtained during the calculation of the ROC curve. The reported scores are based on the F-Measure defined as $F_\alpha = \frac{(1+\alpha)Precision \times Recall}{\alpha \times Precision + Recall}$. As in previous evaluations [8,14,2], we set $\alpha = 0.3$ to weigh precision more than recall.

While the quantitative evaluation reveals superior results, it is important to note that this happens despite being done on unequal grounds. As discussed in Sec. 2, almost all previous approaches to which we compare use additional processes and biases to improve the raw saliency maps by incorporating object properties [11], shape priors [29], face detection [19], or center bias [35]. Our results so far are intentionally stripped of any such additional computations and yet the proposed contextual computation outperforms the state-of-the-art (despite also using the most naïve similarity measure). As we show in Sec. 5, our results can be improved further by incorporating even simple additional steps.

5 Further Improvement by Segmentation

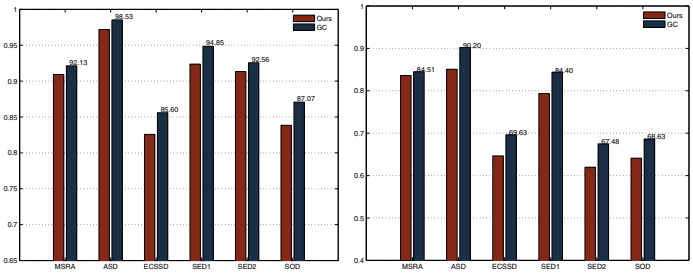
While our raw saliency maps already provide superior results, it is interesting to examine the possible contribution of additional computational steps that are more related to visual objects. To this end, we follow Cheng et al. [14] and use our saliency maps to initialize the GrabCut segmentation algorithm (instead of the manual initialization with a rectangular region, as in the original GrabCut). Unlike Cheng et al. [14], who initialized GrabCut with binary saliency maps based on a fixed threshold, we sought a way to compare results across thresholds so they can be evaluated against the results presented in Sec. 4. Hence, the task becomes one of combining GrabCut with information from our raw (and graded) saliency maps in order to improve overall saliency results.

A possible approach to pursue the above would initialize GrabCut with binarized saliency maps based on all threshold values $0 \leq \tau_i \leq 1$. New foreground regions suggested by GrabCut at each threshold (if they indeed emerge) would then be assigned saliency values in a revised map. This still leaves open the particular strategy of assigning saliency values to aggregated foreground regions. As the segmentation may not capture the entire object or it might include non object regions, careless assignment of saliency values may significantly reduce true-positives (TP) or increase false-positives (FP) and thus reduce performance rather than improving it.

If new foreground regions were assigned their raw saliency values, then FP rate in the revised map could not exceed that in the raw map. Indeed, empirical results based on this approach reduced preformance, implying that the GrabCut segmentation misses parts of the objects that contributed to the results (hence decreasing TP rate). In order to enhance the saliency of foreground regions while preserving the saliency of missed objects parts, we use the following strategy (demonstrated in Fig.7). At each threshold, any suggested foreground region in the revised map is assigned its raw saliency, normalized to the range between the average and maximum values of that region. Only after all threshold values are considered, the remaining regions in the revised map (possibly including missed object parts) are assigned their raw saliency values.

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

Fig. 7. A schematic demonstration of the GrabCut based improvement. The original saliency map (a) is thresholded at different levels (b) to initialize GrabCut, which may suggest new foreground regions at each level (c). New regions are accumulated in the revised map (d). Whenever a region is added to that map, its saliency values are normalized to the range between the average and the maximum values of that region in the original map. The remaining regions are assigned their original saliency values (e).



(a) (b)

Fig. 8. Improvement of our original results using the GrabCut segmentation algorithm. Scores are presented in terms of AUC (panel a) and in terms of F-Measure (panel b).

Using the procedure above, Fig. 8 shows the improvement with respect to our previous results (based on the same evaluation metrics). More specifically, using this segmentation step, original AUC scores improve by $\sim 1\% - 3\%$ and F-measures increase by $\sim 1\% - 5\%$. Since many of the previous algorithms also use additional computations beyond raw saliency, an equal ground comparison to the prior art should consider *these* numbers (rather than those from Sec. 4, which already outperform existing approaches), that indicate that our algorithm exhibits performance which is better than the state-of-the-art by a large margin.

Finally, although it is important to consider objective quantitative measures and results as above, we believe that much of the strength of our approach is revealed at the qualitative level. Indeed, most benchmark databases for saliency detection include relatively simple saliency scenarios, with one (usually visually coherent) salient object typically at a central position. As we argue, the principles underlying previous saliency algorithms (i.e., contrast-based or rarity-based) permit to handle these cases to some extent, but constrain the complexity, frequency, and level of abstraction of the detectable salient objects. In focusing on modeling the context only, our approach is more flexible as indeed was demonstrated already in Fig. 1. Another qualitative comparison for novel images that depict more general saliency scenarios is shown in Fig. 9.

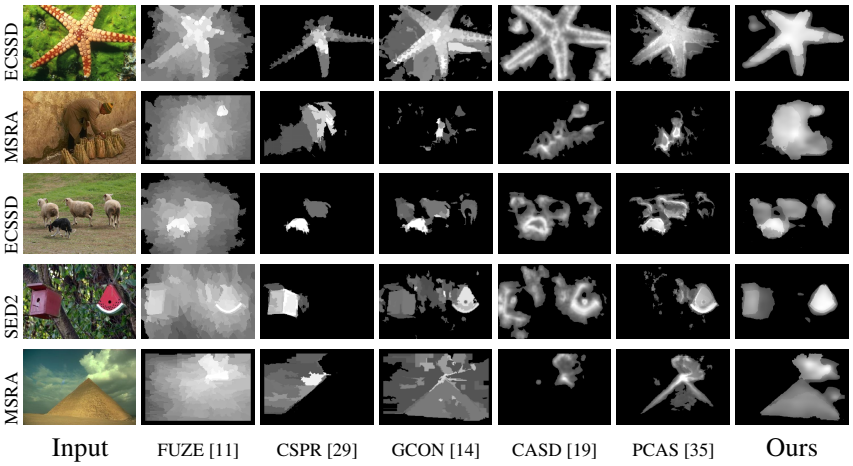


Fig. 9. Example images and normalized saliency maps (thresholded at 30%). The datasets from which the input images are taken are noted on the left. Our saliency maps seem to coherently indicate the saliency of large and complex objects as a whole (first two rows) and allow the detection of multiple salient objects (following two rows). In contrast, no certain level of saliency seems to allow similar detection accuracy by state-of-the-art methods. The last image of a pyramid demonstrates the significance of the enclosure property of visual context for the detection of abstract salient regions.

6 An Unavoidable Commentary about Salient Object Databases

The evaluation of any apcroach inherently depends on two aspects of the dataset to which it is applied. One aspect is ground truth representation. With respect to the datasets above, an apparent problem in this regard is the bounding-box approach used for labeling the MSRA dataset which, as already criticized by Achanta et al. [2], provides limited accuracy. A simple case where this approach may clearly distort evaluation results is when the area ratio between the object and its bounding box is small (e.g., a boomerang). In such a case, false positives within the bounding-box would wrongfully enhance performance while a perfect detection would result in a lower score. To provide a more accurate representation of ground truth, Achanta et al. [2] proposed the ASD dataset in which objects are manually segmented. However, since the data were labeled by a single annotator, the ground truth saliency maps are binary (as is also the case for the ECSSD dataset) whereas the evaluated algorithms may produce graded saliency maps. This discrepancy alone already questions the evaluation reliability.

A second aspect concerns the visual content of the datasets. Although widely used and having size and stimulus variety, the existing datasets are rather restricted in many other ways. For example, as analyzed by Borji et al. [8], these datasets have a strong location-bias and most scenes have low-clutter. An undesired implication is the overfitting of models to existing datasets. Moreover, the suggested ground truth does not allow to evaluate other levels of saliency. This is demonstrated in Borji's benchmark, where methods aiming at fixation prediction show significantly lower performance than methods that seek saliency at the object level.

7 Discussion and Future Directions

We argue that the implicit assumption of having a certain visual constituent at hand when its saliency is measured is at the basis of using different types of context to detect salient objects. The intent for this constituent to be an object motivates its modelling in terms of contrast and rarity. Thus, the nature of visual objects that may be captured is constrained, which necessitates object-specific information and additional computational processes to facilitate better predictions. By modelling visual context instead, we disregard object appearance and reduce these constraints. This allows the saliency of more complex, abstract, or multiple visual objects to emerge. In contrast with previous methods, our approach cannot be categorized as based on contrast or rarity. Our new interpretation of context relies on more basic, general principles.

The ability of our model to outperform the state-of-the-art with no explicit use of object-specific information indicates the dependency of object-based saliency computation on the way context is interpreted in the first place. This is further emphasized by the fact that this superior performance is obtained from low level patches and a single, simple visual feature (i.e., color). Indeed, further development of the suggested theory for contextual emergence of saliency could incorporate additional and more sophisticated features and consider pixels as basic context units. We believe that this would allow to explore the nature of our context based saliency approach for a variety of more complex scenes and perhaps its feasibility for predicting human fixations. However, according to the critisism in section 6, this would require to extend the datasets with more general scenes in terms of complexity, multiplicity, and spatial location. In addition, it would require a new type and more general ground truth that allows to evaluate saliency detection across different levels (fixations and objects). We hope that our novel definition of low-level, non-semantic visual context and the contextual emergence of saliency that follows it would motivate further work in these directions.

Acknowledgements. This work was supported in part by the National Institute for Psychobiology in Israel (grant no. 9-2012/2013) founded by the Charles E. Smith Family, by the Israel Science Foundation (ISF grants no. 259/12 and 1274/11), and by the European Commission in the 7th Framework Programme (CROPS GA no. 246252). We also thank the Frankel Fund, the ABC Robotics initiative, and the Zlotowski Center for Neuroscience at Ben-Gurion University for their generous support.

References

1. Achanta, R., Estrada, F.J., Wils, P., Süsstrunk, S.: Salient region detection and segmentation. In: Gasteratos, A., Vincze, M., Tsotsos, J.K. (eds.) ICVS 2008. LNCS, vol. 5008, pp. 66–75. Springer, Heidelberg (2008)
2. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1597–1604 (2009)
3. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels. École Polytechnique Fédéral de Lausssanne (EPFL), Tech. Rep. (2010)

4. Achanta, R., Susstrunk, S.: Saliency detection for content-aware image rresizing. In: Proceedings of the IEEE International Conference on Image Processing, pp. 1005–1008 (2009)
5. Bar, M., Ullman, S.: Spatial context in recognition. Perception 25, 343–352 (1996)
6. Biederman, I.: Perceiving real-world scenes. Science 177, 77–80 (1972)
7. Biedermanl, I., Mezzanote, R., Rabinowitx, J.: Scene perception: Detecting and judging objects undergoing relational violations. Cognitive Psychology 14(2), 143–177 (1982)
8. Borji, A., Sihite, D.N., Itti, L.: Salient object detection: A benchmark. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 414–429. Springer, Heidelberg (2012)
9. Botev, Z., Grotowski, J., Kroese, D.: Kernel density estimation via diffusion. The Annals of Statistics 38(5), 2916–2957 (2010)
10. Bruce, N., Tsotsos, J.: Saliency based on information maximization. In: Neural Information Processing Systems, pp. 155–162 (2005)
11. Chang, K., Liu, T., Chen, H., Lai, S.: Fusing generic objectness and visual saliency for salient object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 914–921 (2011)
12. Chen, L.Q., Xie, X., Fan, X., Ma, W.Y., Zhang, H.J., Zhou, H.Q.: A visual attention model for adapting images on small displays. Multimedia Systems 9(4), 353–364 (2003)
13. Chen, T., Cheng, M.M., Tan, P., Shamir, A., Hu, S.M.: Sketch2photo: internet image montage. ACM Transactions on Graphics (TOG) 28, 124 (2009)
14. Cheng, M., Zhang, G., Mitra, N., Huang, X., Hu, S.: Global contrast based salient region detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 409–416 (2011)
15. Christopoulos, C., Skodras, A., Ebrahimi, T.: The jpeg2000 still image coding system: an overview. IEEE Transactions on Consumer Electronics 46(4), 1103–1127 (2000)
16. Duncan, J., Humphreys, G.: Visual search and stimulus similarity. American Psychological Association 96(3), 433–458 (1989)
17. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge (VOC2012) Results (2012), `http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html`
18. Goferman, S., Tal, A., Zelnik-Manor, L.: Puzzle-like collage. Computer Graphics Forum 29, 459–468 (2010)
19. Goferman, S., Zelnik-Manor, L., Tal, A.: Context-aware saliency detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(10), 1915–1926 (2012)
20. Guo, C., Zhang, L.: A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. IEEE Transactions on Image Processing 19(1), 185–198 (2010)
21. Han, J., Ngan, K., Li, M., Zhang, H.J.: Unsupervised extraction of visual attention objects in color images. IEEE Transactions on Circuits and Systems for Video Technology 16(1), 141–145 (2006)
22. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Neural Information Processing Systems, pp. 545–552 (2006)
23. Hou, X., Zhang, L.: Saliency detection: A spectral residual approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
24. Hu, Y., Xie, X., Ma, W.-Y., Chia, L.-T., Rajan, D.: Salient region detection using weighted feature maps based on the human visual attention model. In: Aizawa, K., Nakamura, Y., Satoh, S. (eds.) PCM 2004. LNCS, vol. 3332, pp. 993–1000. Springer, Heidelberg (2004)
25. Itti, L., Baldi, P.: Bayesian surprise attracts human attention. In: Neural Information Processing Systems, pp. 547–554 (2005)

26. Itti, L., Koch, C.: Computational modelling of visual attention 2(3), 194–203 (2001)
27. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11), 1254–1259 (1998)
28. Jian, L., Saliency, L.M.A.X.H.H.: detection based on frequency and spatial domain analyses. In: British Machine Vision Conference, pp. 86.1–86.11 (2011)
29. Jiang, H., Wang, J., Yuan, Z., Liu, T., Zheng, N., Li, S.: Automatic salient object segmentation based on context and shape prior. In: British Machine Vision Conference (2011)
30. Ko, B.C., Nam, J.Y.: Object-of-interest image segmentation based on human attention and semantic region clustering. J. Opt. Soc. Am. A 23(10), 2462–2470 (2006)
31. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry 4(4), 219–227 (1985)
32. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., Shum, H.: Learning to detect a salient object. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(2), 353–367 (2011)
33. Ma, Y., Zhang, H.: Contrast-based image attention analysis by using fuzzy growing. In: Proceedings of the Eleventh ACM international conference on Multimedia, pp. 374–381 (2003)
34. Marchesotti, L., Cifarelli, C., Csurka, G.: A framework for visual saliency detection with applications to image thumbnailing. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2232–2239 (2009)
35. Margolin, R., Tal, A., Zelnik-Manor, L.: What makes a patch distinct? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2013)
36. 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: Proceedings of the IEEE International Conference on Computer Vision, pp. 416–423 (2001)
37. Navon, D.: Forest before the trees: the precedence of global features in visual perception. Cognitive Psychology 9, 353–383 (1977)
38. Nothdurft, H.: Salience from feature contrast: additivity across dimensions. Vision Research 40(10), 1073–1078 (2000)
39. Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 490–503. Springer, Heidelberg (2006)
40. Oliva, A., Torralba, A., Castelhano, M.S., Henderson, J.: Top-down control of visual attention in object detection. In: Proceedings of the IEEE International Conference on Image Processing (2003)
41. Palmer, S.: The effects of contextual scenes on the identification of objects. Memory & Cognition 3, 519–526 (1975)
42. Rother, C., Bordeaux, L., Hamadi, Y., Blake, A.: Autocollage. In: ACM SIGGRAPH 2006 Papers. pp. 847–852. ACM Press (2006)
43. Rutishauser, U., Walther, D., Koch, C., Perona, P.: Is bottom-up attention useful for object recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 37–44 (2004)
44. Silverman, B.: Density estimation for statistics and data analysis, vol. 26. CRC press (1986)
45. Suh, B., Ling, H., Bederson, B.B., Jacobs, D.W.: Automatic thumbnail cropping and its effectiveness. In: Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology, pp. 95–104 (2003)
46. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. Cognitive Psychology 12(1), 97–136 (1980)
47. Wang, J., Quan, L., Sun, J., Tang, X., Shum, H.Y.: Picture collage. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 347–354 (2006)

48. Wang, L., Xue, J., Zheng, N., Hua, G.: Automatic salient object extraction with contextual cue. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 105–112 (2011)
49. Wang, Y.S., Tai, C.L., Sorkine, O., Lee, T.Y.: Optimized scale-and-stretch for image resizing. ACM Transactions on Graphics (TOG) 27, 118 (2008)
50. Xue, J., Li, C., Zheng, N.: Proto-object based rate control for jpeg2000: an approach to content-based scalability 20(4), 1177–1184 (2011)
51. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1155–1162. IEEE (2013)
52. Zhai, Y., Shah, M.: Visual attention detection in video sequences using spatiotemporal cues. In: Proceedings of the 14th Annual ACM International Conference on Multimedia, pp. 815–824