Context-Aware Saliency Detection

Stas Goferman, Lihi Zelnik-Manor, *Member*, *IEEE*, and Ayellet Tal

Abstract—We propose a new type of saliency—context-aware saliency—which aims at detecting the image regions that represent the scene. This definition differs from previous definitions whose goal is to either identify fixation points or detect the dominant object. In accordance with our saliency definition, we present a detection algorithm which is based on four principles observed in the psychological literature. The benefits of the proposed approach are evaluated in two applications where the context of the dominant objects is just as essential as the objects themselves. In image retargeting, we demonstrate that using our saliency prevents distortions in the important regions. In summarization, we show that our saliency helps to produce compact, appealing, and informative summaries.

Index Terms—Image saliency, visual saliency, context aware.

1 Introduction

Please take a look at the images on the top row of Fig. 1. How would you describe them? Probably you'd say "a smiling girl," "a figure in a yellow flower field," and "a weightlifter in the Olympic games" (or something similar).[1] Each title describes the essence of the corresponding image—what most people think is important or *salient*.

A profound challenge in computer vision is the detection of the salient regions of an image. The numerous applications (e.g., [4], [32], [27], [30], [9]) that make use of these regions have led to different definitions and interesting detection algorithms. Classically, algorithms for saliency detection focused on identifying the fixation points that a human viewer would focus on at first glance [16], [15], [35], [6], [13], [20]. This type of saliency is important for understanding human attention, as well as for specific applications such as autofocusing. Others have concentrated on detecting a single dominant object of an image [21], [14], [12]. For instance, in Fig. 1, such methods aim to extract the "girl," the "figure," and the "athlete" (third row). This type of saliency is useful for several high-level tasks, such as object recognition [30] or segmentation [28].

There are, however, applications where the context of the dominant objects is just as essential as the objects themselves. Examples include image classification [23], summarization of a photo collection [27], thumbnailing [32], and retargeting [29]. For these applications, the detected regions in Fig. 1 should correspond to the titles you gave above. The regions on the bottom row of Fig. 1 match these titles better than the regions on the third row.

This calls for introducing a new type of saliency—context-aware saliency. Here, the goal is to identify the pixels that correspond to the bottom row (and to the titles). According to this concept, the salient regions should contain not only the prominent objects but also the parts of the background that convey the context.

We differentiate between three types of images, as illustrated in Fig. 1. In the girl's case, the background is not interesting; hence, we expect the extracted salient region to coincide with the salient object. In the flower-field's case, the texture of the flowers is essential for understanding the content. However, only a small portion of it—the portion surrounding the figure—suffices. In the weightlifter's case, the weights and the Olympic logo are vital for conveying the scene. This is not necessarily the portion surrounding the athlete, but rather a unique part of the background. Therefore, detecting the prominent object together with naive addition of its immediate surrounding will not suffice.

This paper proposes a novel algorithm for context-aware saliency detection. The underlying idea is that salient regions are distinctive with respect to both their local and global surroundings. Hence, the unique parts of the background, and not only the dominant objects, would be marked salient by our algorithm (e.g., the Olympics logo in Fig. 1). Moreover, to comply with the Gestalt laws, we prioritize regions close to the foci of attention. This maintains the background texture when it is interesting, such as in the case of the flower field in Fig. 1.

We demonstrate the utility of our context-aware saliency in retargeting [4], [29], [24]. We show that our saliency can successfully mark the regions that should be kept untouched.

The contribution of this paper is hence threefold. First, we introduce principles for context-aware saliency (Section 3). Second, we propose an algorithm that detects this saliency (Section 4) and present results on images of various types (Section 5). Last but not least, we demonstrate the applicability of our saliency (Section 7).

A preliminary version of this work appeared in [10].

1. These descriptions were obtained by collecting titles given by 12 different people. See samples in the second row of Fig. 1.

- *S. Goferman can be reached at 37 Raoul Wallenberg Street, Haifa 34990, Israel. E-mail: stasix@gmail.com.*
- *L. Zelnik-Manor and A. Tal are with the Department of Electrical Engineering, Technion-Israel Institute of Technology, Haifa 32000, Israel. E-mail: {lihi, ayellet}@ee.technion.ac.il.*

Manuscript received 11 Jan. 2011; revised 18 Oct. 2011; accepted 15 Dec. 2011; published online 22 Dec. 2011.
Recommended for acceptance by C. Rother.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2011-01-0026.
Digital Object Identifier no. 10.1109/TPAMI.2011.272.

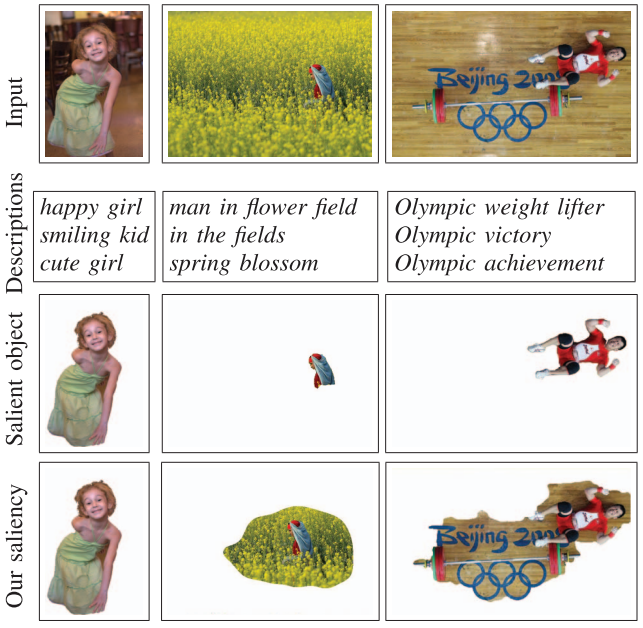


Fig. 1. Our context-aware saliency results (bottom) comply with the descriptions that people provided (samples in the second row) for the input images (top). People tend to describe the *scene* rather than the *dominant object*. Extracting the single most dominant object, as shown on the third row, might miss the essence of the scene. Conversely, we maintain all the essential regions of the image.

2 Related Work

Many visual attention approaches have been proposed for detecting a few key locations with maximum local saliency and employ biologically motivated low-level features [16], [15], [22], [35], [6], [13], [20]. They are driven by low-level stimulus such as intensity, color, orientation, and texture. While these approaches manage to find a few fixation points, they do not aim at finding regions of visual attention. For example, the saliency results of the motorcycle picture in Fig. 2b illustrate why this is insufficient: The high local contrast includes all the transitions between the background stripes.

Other approaches focus on global features, e.g., [1], [14], [12]. These methods are based on finding regions in the image which imply unique frequencies in the Fourier domain. Therefore, they are able to quickly locate visual "pop-outs" that can serve as candidates for salient objects. These methods are very fast, but since they are based on global considerations, they do not detect object boundaries accurately, as illustrated in Fig. 2c.

The approach we propose unifies local and global saliency by measuring the similarity between each image patch and other image patches, both locally and globally. In [31], the resemblance between patches was also used as a cue for saliency estimation; however, there only local neighborhoods were compared. In [5], an image region is considered salient if it cannot be explained by anything similar in other portions of the image. Since their goal was to detect salient regions, they search for ensembles of patches that are unique globally. The lack of local considerations leads to coarse detection results.

An attempt to incorporate local and global features for the purpose of detecting a single salient object is proposed in [21]. This is done by using three features: a center-surround histogram, color spatial distribution features, and local multiscale contrast features. While this method computes saliency maps, which have been shown to be useful for extracting rectangular bounding boxes of a single object of interest, they are less appropriate for extracting accurate regions-of-interest, the context of the salient object, and sceneries or crowds.

The methods that use global features consider one aspect of the context—the saliency of an object depends on the interrelated conditions in which it exists. Like them, we consider global uniqueness as a key ingredient of saliency. However, our goal is somewhat different—we wish to extract the salient objects together with the parts of the discourse that surrounds them and can shed light on the meaning of the image. To achieve this, we propose a method that is inspired by psychological evidence and combine both local and global considerations. As illustrated in the motorcyclist of Fig. 2, this allows us to nicely detect the motorcyclist and his reflection since the motorcyclist is distinctive both locally and globally and the reflection provides the context.

3 Principles of Context-Aware Saliency

Our context-aware saliency follows four basic principles of human visual attention, which are supported by psychological evidence [33], [36], [18], [19]:

1. Local low-level considerations, including factors such as contrast and color.
2. Global considerations which suppress frequently occurring features while maintaining features that deviate from the norm.
3. Visual organization rules which state that visual forms may possess one or several centers of gravity about which the form is organized.
4. High-level factors such as priors on the salient object location and object detection.

Related work typically follows only some of these principles and hence might not provide the results we

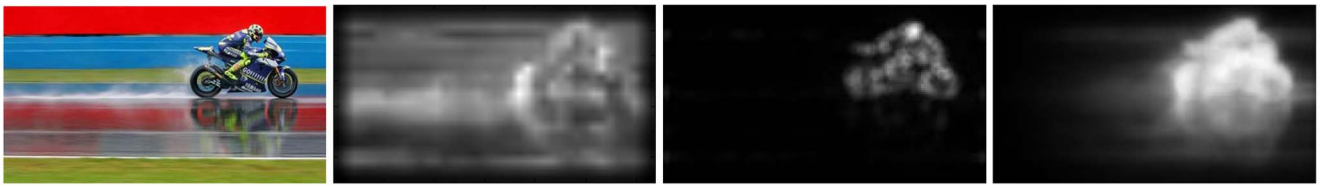


Fig. 2. Comparing different approaches to saliency estimation.

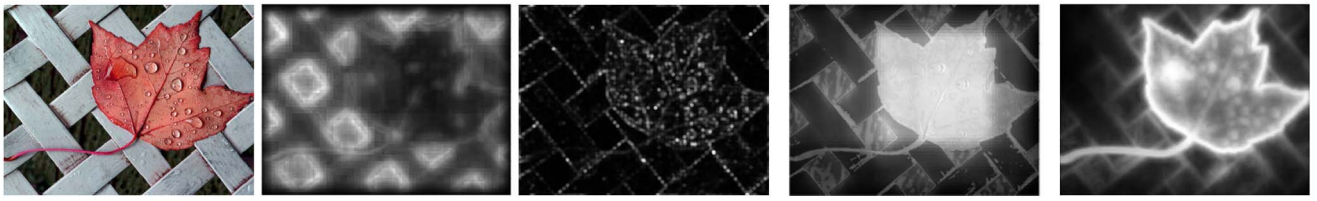


Fig. 3. Comparing different approaches to saliency.

desire. The biologically motivated algorithms for saliency estimation [16], [15], [35], [6], [13], [20] are based on principle 1. Therefore, in Fig. 3b, they detect mostly the intersections on the fence. The approaches of [14], [12] focus on principle 2. Therefore, in Fig. 3c, they detect mostly the drops on the leaf. In [21], an algorithm was proposed for extracting rectangular bounding boxes of a single object of interest. This was achieved by combining local saliency with global image segmentation, thus can be viewed as incorporating principles 1 and 2. In Fig. 3d, they detect as salient both the fence and the leaf, with higher importance assigned to the leaf.

We wish to extract the salient objects together with the parts of the discourse that surrounds them and can shed light on the meaning of the image. To achieve this, we propose a novel method for realizing the four principles. This method defines a novel measure of distinctiveness that combines principles 1, 2, and 3. As illustrated in Fig. 3e, our algorithm detects as salient the leaf, the water-drops, and just enough of the fence to convey the context. Principle 4 is added as postprocessing.

4 Detection of Context-Aware Saliency

In this section, we propose an algorithm for realizing principles 1-4. In accordance with principle 1, areas that have distinctive colors or patterns should obtain high saliency. Conversely, homogeneous or blurred areas should obtain low saliency values. In agreement with principle 2, frequently occurring features should be suppressed. According to principle 3, the salient pixels should be grouped together and not spread all over the image.

This section is structured as follows (Fig. 4). We first define single-scale local-global saliency based on principles 1-3. Then, we further enhance the saliency by using multiple scales. Next, we modify the saliency to further accommodate principle 3. Finally, principle 4 is implemented as post-processing.

4.1 Local-Global Single-Scale Saliency

There are two challenges in defining our saliency. The first is how to define distinctiveness both locally and globally. The second is how to incorporate positional information.

According to principles 1-2, a pixel is salient if its appearance is unique. We should not, however, look at an

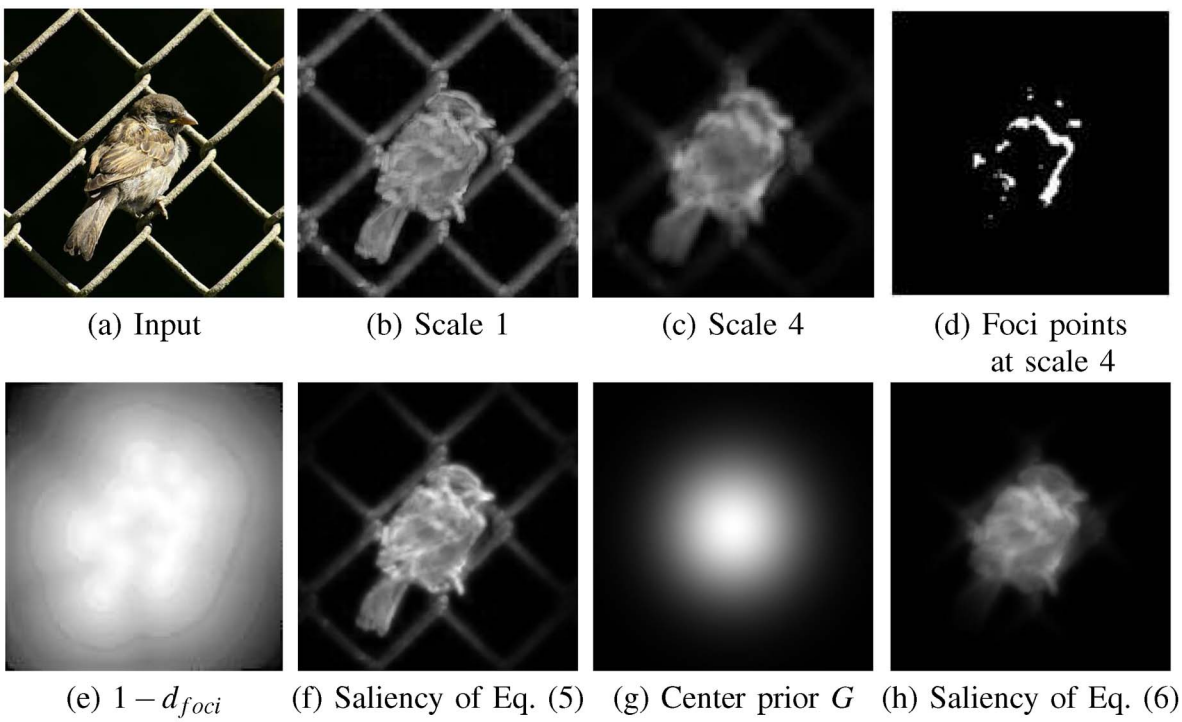


Fig. 4. The steps of our saliency estimation algorithm.

isolated pixel, but rather at its surrounding patch, which gives an immediate context. For now we consider a single patch of scale r at each pixel. Thus, a pixel i is considered salient if the appearance of the patch p_i centered at pixel i is distinctive with respect to all other image patches.

Specifically, let $d_{color}(p_i, p_j)$ be the euclidean distance between the vectorized patches p_i and p_j in CIE L^*a^*b color space, normalized to the range $[0, 1]$. Pixel i is considered salient when $d_{color}(p_i, p_j)$ is high $\forall j$.

In practice, to evaluate a patch's uniqueness, there is no need to incorporate its dissimilarity to all other image patches. It suffices to consider the K most similar patches (if the most similar patches are highly different from p_i, then clearly all image patches are highly different from p_i). Hence, for every patch p_i, we search for the K most similar patches $\{q_k\}_{k=1}^{K}$ in the image, according to $d_{color}(p_i, p_j)$. A pixel i is salient when $d_{color}(p_i, q_k)$ is high $\forall k \in [1, K]$ ($K = 64$ in all our experiments).

According to principle 3, the positional distance between patches is also an important factor. Background patches are likely to have many similar patches both near and far-away in the image. This is in contrast to salient patches which tend to be grouped together. This implies that a patch p_i is salient when the patches similar to it are nearby, and it is less salient when the resembling patches are far away.

Let $d_{position}(p_i, q_k)$ be the euclidean distance between the positions of patches p_i and q_k, normalized by the larger image dimension. Based on the observations above, we define a dissimilarity measure between a pair of patches as

$$d(p_i, q_k) = \frac{d_{color}(p_i, q_k)}{1 + c \cdot d_{position}(p_i, q_k)}, \tag{1}$$

where $c = 3$ in our implementation. This dissimilarity measure is proportional to the difference in appearance and inversely proportional to the positional distance.

As mentioned above, a pixel i is salient when $d(p_i, q_k)$ is high $\forall k \in [1, K]$. Hence, the single-scale saliency value of pixel i at scale r is defined as

$$S_i^r = 1 - exp\left\{-\frac{1}{K}\sum_{k=1}^{K} d\left(p_i^r, q_k^r\right)\right\}. \tag{2}$$

4.2 Multiscale Saliency Enhancement

Background pixels (patches) are likely to have similar patches at multiple scales, e.g., in large homogeneous or blurred regions. This is in contrast to more salient pixels that could have similar patches at a few scales but not at all of them. Therefore, we incorporate multiple scales to further decrease the saliency of background pixels, improving the contrast between salient and nonsalient regions.

Multiple scales are utilized by representing each pixel by the set of multiscale image patches centered at it. A pixel is considered salient if it is consistently different from other pixels in multiple scales. One way to compute such global saliency is to consider a pixel to be salient if its multiscale K-most similar patches are different from it.

For a patch p_i of scale r, we consider as candidate neighbors all the patches in the image whose scales are $R_q = \{r, \frac{1}{2}r, \frac{1}{4}r\}$. Among all these patches, the K most similar patches according to (1) are found and used for computing the saliency. Hence, (2) can be rewritten as (where $r_k \in R_q$)

$$S_i^r = \left[1 - exp\left\{-\frac{1}{K}\sum_{k=1}^{K} d\left(p_i^r, q_k^{r_k}\right)\right\}\right]. \tag{3}$$

The saliency map S_i^r at each scale is normalized to the range $[0, 1]$ and interpolated back to original image size.

Furthermore, we represent each pixel by the set of multiscale image patches centered at it. Let $R = \{r_1, \ldots, r_M\}$ denote the set of patch sizes to be considered for pixel i. The saliency at pixel i is taken as the mean of its saliency at different scales:

$$\bar{S}_i = \frac{1}{M}\sum_{r \in R} S_i^r, \tag{4}$$

where S_i^r is defined in (3). The larger $\bar{S}_i$ is, the more salient pixel i is and the larger is its dissimilarity (in various levels) to the other patches.

In our implementation, we scale all the images to the same size of 250 pixels (largest dimension). We compute the saliency for all image pixels; however, when searching for the nearest neighbors we only consider patches of size 7×7 with 50 percent overlap. We use four scales: $R = \{100\%, 80\%, 50\%, 30\%\}$. The smallest scale allowed in R_q is 20 percent of the original image scale.

Figs. 4b and 4c demonstrate the difference between the saliency maps obtained at different scales. While the fine-scale result detects all the details, including those of the background, the coarse scale result detects mostly the bird.

4.3 Including the Immediate Context

According to Gestalt laws, visual forms may possess one or several centers of gravity about which the form is organized [19] (principle 3). This suggests that areas that are close to the foci of attention should be explored significantly more than far-away regions. When the regions surrounding the foci convey the context, they draw our attention and thus are salient.

We simulate this visual contextual effect in two steps. First, the most attended localized areas at each scale are extracted from the saliency maps produced by (3). A pixel is considered attended at scale r if its saliency value exceeds a certain threshold ($S_i^r > 0.8$ in the examples shown in this paper), see Fig. 4d.

Then, each pixel outside the attended areas is weighted according to its euclidean distance to the closest attended pixel. Let $d_{foci}^r(i)$ be the euclidean positional distance between pixel i and the closest focus of attention pixel at scale r, normalized to the range $[0, 1]$ (Fig. 4e). The saliency of pixel i is redefined as

$$\hat{S}_i = \frac{1}{M}\sum_{r \in R} S_i^r\left(1 - d_{foci}^r(i)\right). \tag{5}$$

Note that the saliency of noninteresting regions, such as blurred or homogeneous regions, remains low since $\bar{S}$ of (4) will dominate. However, the saliency of interesting background in the neighborhood of the salient objects will be increased by (5). This explains why in Fig. 1 parts of the flower field were detected as salient in the center example, whereas the girl on the left was segmented accurately. In Fig. 4f, this last step enhances the bird and attenuates the far parts of the background wire.

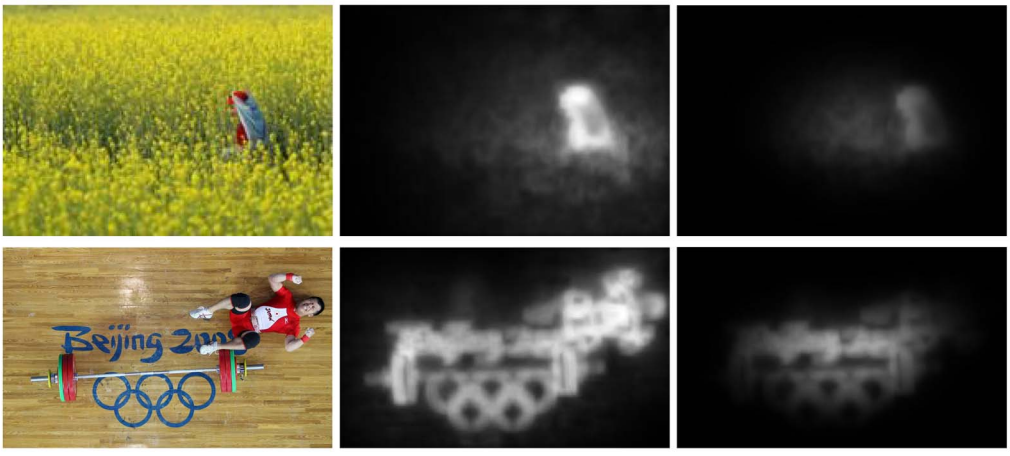


Fig. 5. Comparing saliency results (b) without and (c) with the center prior.

4.4 Center Prior

It is a well-known fact that when humans take pictures they often frame the objects of interest near the center of the image. As a consequence it was shown in [17] that a saliency map based on the distance of each pixel to the center of the image provides a better prediction of the salient object than many previous saliency models. Inspired by this, we further incorporate a center prior to our saliency estimation.

Let $G(\boldsymbol{\sigma}_x, \boldsymbol{\sigma}_y)$ be a 2D Gaussian positioned at the center of the image Fig. 4g. The horizontal variance is set to $\sigma_x = \frac{\#\text{columns}}{6}$ and the vertical variance to $\boldsymbol{\sigma}_y = \frac{\#\text{rows}}{6}$. Our final saliency (see Fig. 4h) of a pixel is defined as

$$S_i = \hat{S}_i G_i, \tag{6}$$

where G_i is the value of pixel i in the map G.

Fig. 5 provides a visual comparison between the saliency maps obtained with and without the center prior. Interestingly, the saliency maps of (5), which do not include the center prior, seem more visually appealing than the saliency maps of (6) which include the center prior. Nevertheless, later on in Section 5, we provide quantitative evaluation on ground-truth data, which shows that more accurate results are obtained when the center prior is included.

4.5 High-Level Factors

Finally, the saliency map could be further enhanced using high-level factors, such as recognized objects or face detection. For example, one could incorporate the face detection algorithm of [34], which generates 1 for face pixels and 0 otherwise. The saliency map of (6) can then be modified by taking the maximum value of the saliency map and the face map. We view this step as an a posteriori refinement of the saliency map and hence exclude it from our experiments.

5 Empirical Evaluation

To evaluate the quality of the proposed approach, we provide in this section both qualitative as well as quantitative evaluation. As was shown in Fig. 5, in many cases the saliency maps of (5) which do not include the center prior look more convincing visually than the saliency maps of (6), which do include the center prior. Hence, all the qualitative results presented were obtained using the saliency of (5), i.e., excluding the center prior. Below, we present quantitative evaluation on standard benchmarks both with and without the center prior. Later on, we discuss what leads to the difference between the two evaluation options.

5.1 Qualitative Evaluation

Figs. 6, 7, and 8 compare our results with the biologically inspired local-contrast approach of [35] and the spectral residual global approach of [14]. Later on, in Fig. 12, we compare our results with the single-object detection of [21].

As will be shown next, the method of [35] detects as salient many noninteresting background pixels since it does not consider any global features. The approach of [14] fails to detect many pixels on the prominent objects since it does not incorporate local saliency. Our approach consistently detects with higher accuracy the pixels on the dominant objects and their contextual surroundings.

We distinguish between three cases. The first case (Fig. 6) includes images that show a single salient object over an uninteresting background. For such images, we expect that only the object's pixels will be identified as salient. In [35], some pixels on the objects are very salient, while other pixels—both on the object and on the background—are partially salient as well. In [14], the background is nicely excluded; however, many pixels on the salient objects are not detected as salient. Our algorithm manages to detect the pixels on the salient objects and only them.

The second case includes images where the immediate surroundings of the salient object shed light on the story the image tells. In other words, the surroundings are also salient. Unlike the other approaches, our results capture the salient parts of the background, which convey the context. For example, in Fig. 7, the swimmer is detected together with the foam he generates, and in Fig. 2 the motorcyclist is detected together with his reflection and part of the race track.

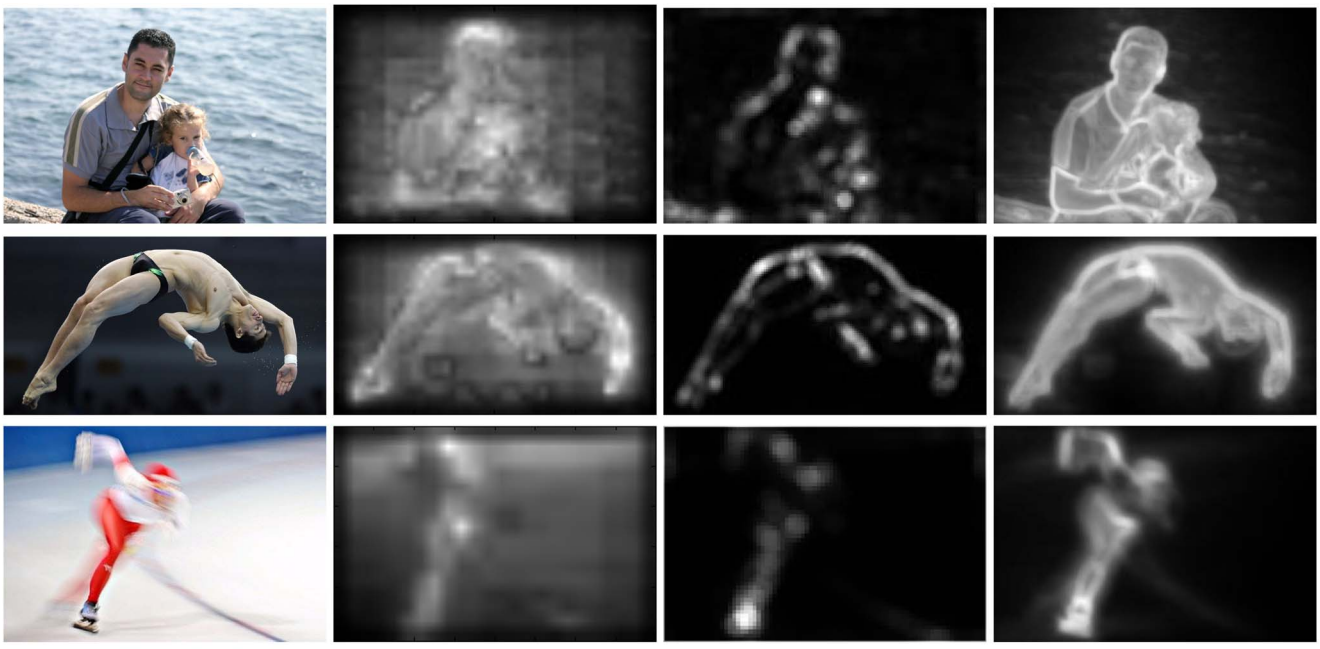


Fig. 6. Comparing saliency results on images of a single object over an uninteresting background.

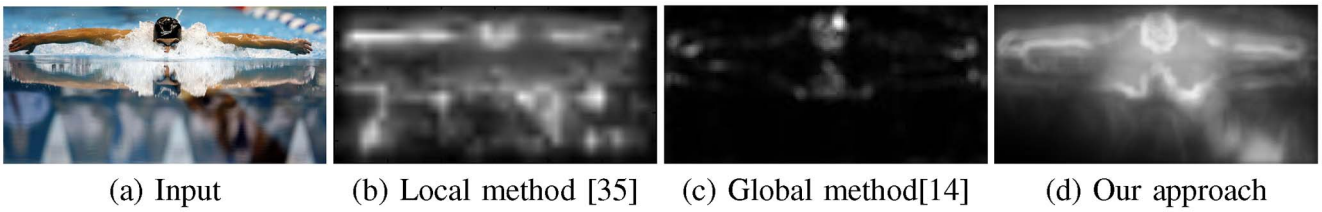


Fig. 7. Comparing saliency results on images in which the immediate surroundings of the salient object are also salient.

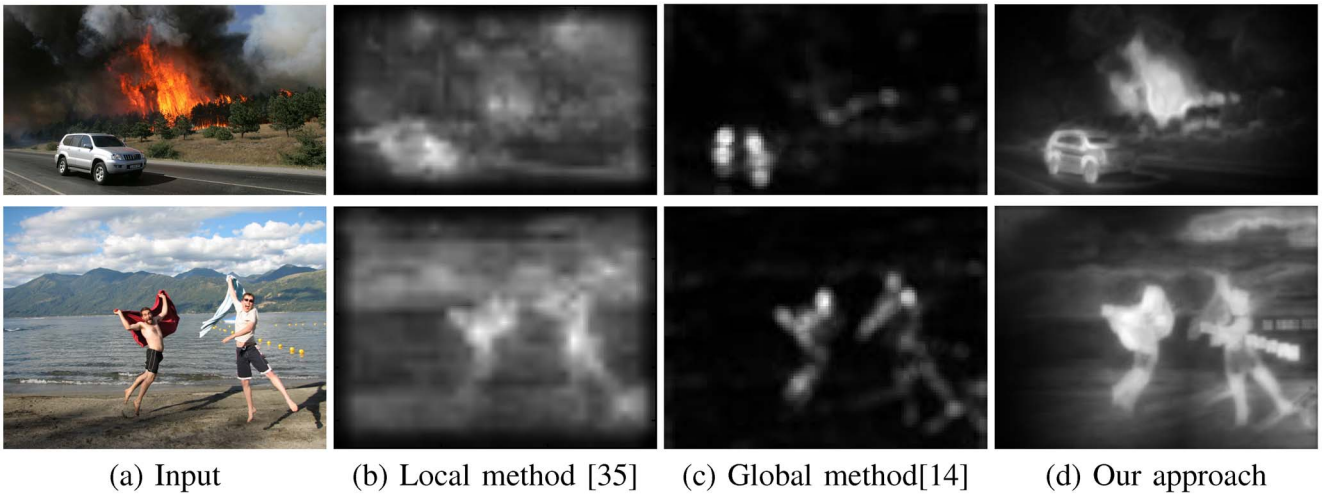
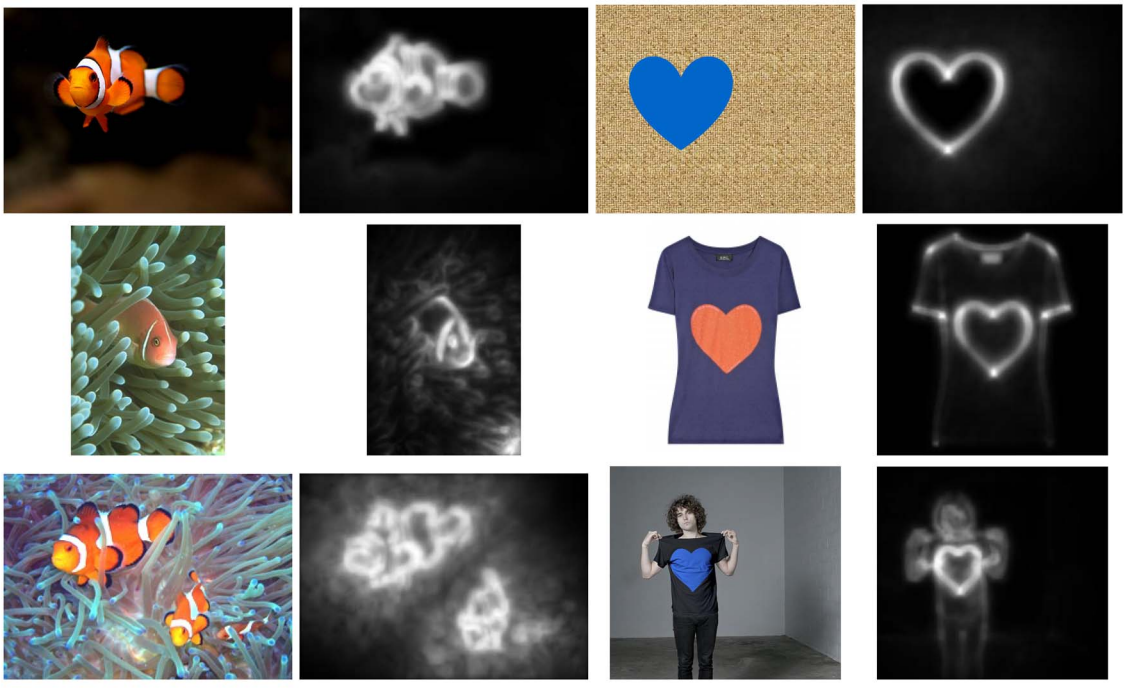


Fig. 8. Comparing saliency results on images of complex scenes.

The third case includes images of complex scenes. For instance, Fig. 8 shows an image of a car in a fire scene and an image of two cheering guys by the lake and mountains. It can be observed that our approach detects as salient both the vehicle and the fire in the first scene and the guys with part of the scenery in the other one.

Fig. 9 demonstrates how our approach captures context. We detect the saliency of a certain object on varying backgrounds. It can be seen that when the background has no importance, only the dominant object is detected, whereas if the background is meaningful, parts of it are included as well.

Fig. 10 compares our approach with the irregularity detection of [5]. Recall that the goals differ since they look for an ensemble of patches that cannot be explained by other image regions. Yet, it can be seen that our saliency



(a) Various scenes of clown fish (b) Various scenes of hearts

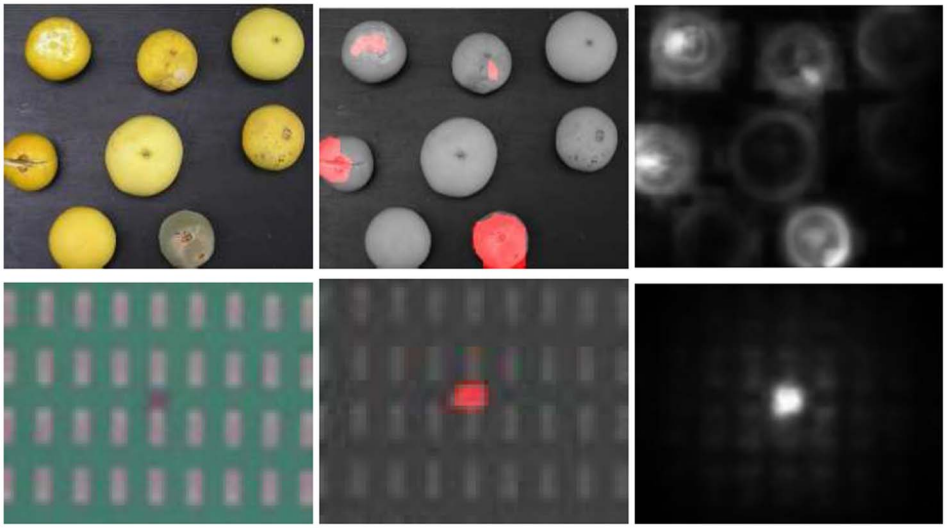
Fig. 9. (a) The left column shows images of clown fish. On the top, the background is uniform and our algorithm detects only the fish as salient. In the middle, the background is interesting, so part of it is captured. On the bottom, the two fish are captured, as well as some of their surrounding background. (b) Similar behavior is shown on the right, where only the heart is detected when the background is a texture, the edges of the shirt are also detected in the middle image, and finally the man's facial features and arms are detected on the bottom.

maps emphasize the irregular patterns more accurately and, unlike [5], provides some of the contextual pattern.

5.2 Quantitative Evaluation

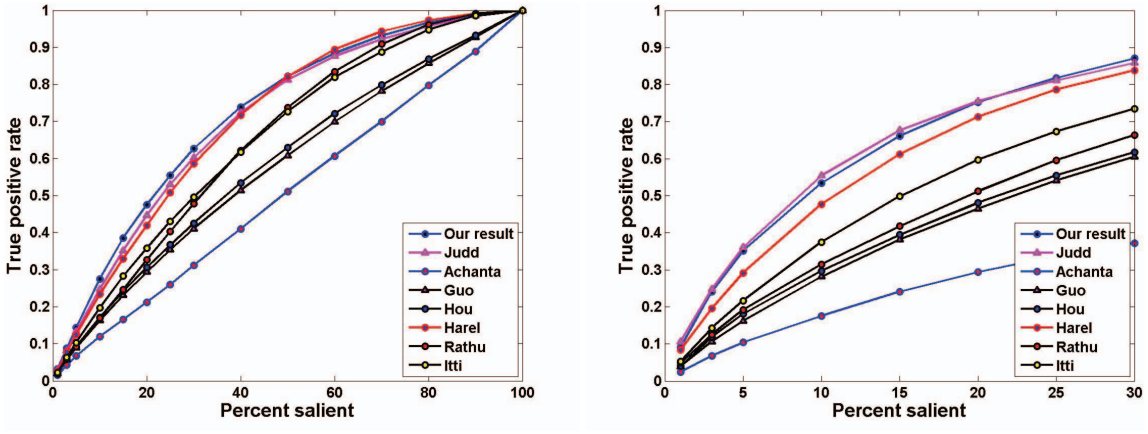
To obtain a quantitative evaluation, we compare ROC curves on two different benchmarks. Results are presented for seven different state-of-the-art algorithms [1], [12], [13], [14], [16], [17], [26]. In all of our experiments we have *not used face detection*. While object and face detection can clearly improve the results, we wish to evaluate the quality of our model without this a posteriori step.

The first database was presented in [14]. It includes 62 images of different scenes where ground-truth was obtained by asking people to "select regions where objects are presented," i.e., the salient regions were marked by hand. In part of the images, only the dominant object was marked, while in others parts of the essential context were also selected. A somewhat different database was presented



(a) Input (b) Boiman's result (c) Our saliency

Fig. 10. Comparison of our saliency to the irregularity detection algorithm of [5]. While they detect roughly the irregular patterns, we detect them more accurately while giving some context.



(a) Database of [14] (b) Database of [17]

Fig. 11. Quantitative evaluation. ROC curves for the databases of [14] (left) and [17] (right). Our results are comparable to [17] even though we do not use any learning or high-level object detection, while [17] uses both.

in [17]. There, images were presented to human observers for 3 seconds each and eye tracking data was collected and averaged. The database includes 100 images for testing. It is important to note that since the images were displayed for 3 seconds only, the eye-tracking data captures mostly what one sees at a first glance.

Fig. 11 shows that our algorithm outperforms almost all algorithms and that our saliency results are comparable to the best results obtained by Judd et al. [17]. This is surprising since their method is extremely involved in comparison to our approach. The approach of [17] is based on a heavy learning phase which combines six different saliency models and three different object detectors (of cars, faces, and people). Our approach, on the other hand, does not require any learning phase and the presented results were obtained without using any high-level information.

Methods like [21] are not designed for such complex scenes, but rather for single dominant-object images. We do not have access to their code; hence we cannot show their results on Figs. 7 and 8. Instead, comparisons are shown on images from their paper (Fig. 12). In [21], a large database of single-object images is presented with impressive extraction results. In the left two images of Fig. 12, they successfully extract the "man" and the "bird." Conversely, our saliency maps indicate that the images show "two men talking" (as both are marked salient) and a "bird on a branch feeding its fledglings," hence providing the context. The image of the woman demonstrates another feature of our algorithm. While Liu et al. [21] detect the upper body of the woman (the black dress is captured due to its salient color), our algorithm marks as salient the entire woman as well as some of the stone wall, thus capturing her posing for the camera.

Our results indicate a gap between qualitative and quantitative evaluation. When visualizing saliency maps, one tends to prefer the results of (5), which do not include the center prior, while quantitatively the results of (6) are better. The difference between the two evaluation schemes stems from their nature. When looking at a saliency map, one expects to see all the pixels on the dominant object highlighted. However, the ground-truth in the database of [17] was obtained by tracking eye-gaze; hence, not all the pixels were marked as salient, but rather only those that were attended frequently enough. In fact, the ground-truth saliency maps are extremely sparse. When comparing our saliency maps with the ground-truth, only these sparse points are considered, and hence visual assessment is somewhat biased.

5.3 The Effect of the Parameters

Finally, for all the experiments described above, the parameters were kept fixed—no user fine-tuning was done. Yet, we want to test the robustness of our algorithm to the parameters and to analyze their effect. In particular, we repeated the experiments, while varying c from (1), K from (2), and R from (4). Fig. 13 shows typical results when

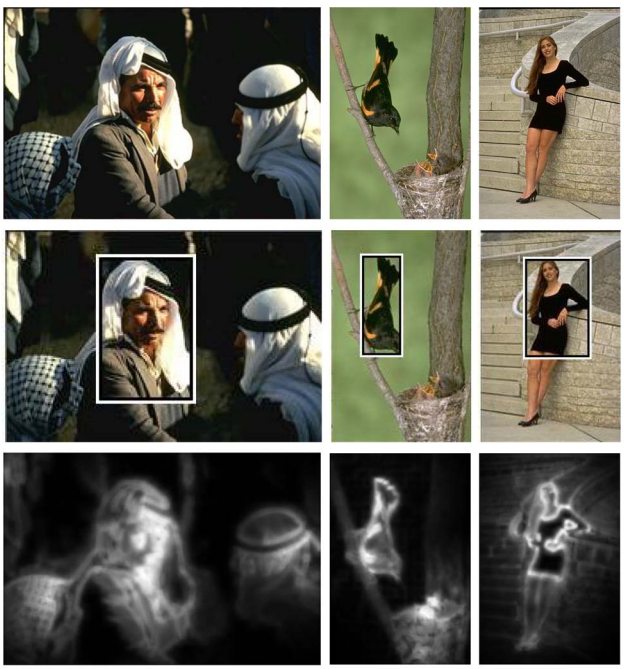


Fig. 12. Comparing our saliency results with [21]. Top: Input images. Middle: The bounding boxes obtained by Liu et al. [21] capture a single main object. Bottom: Our saliency maps convey the story.

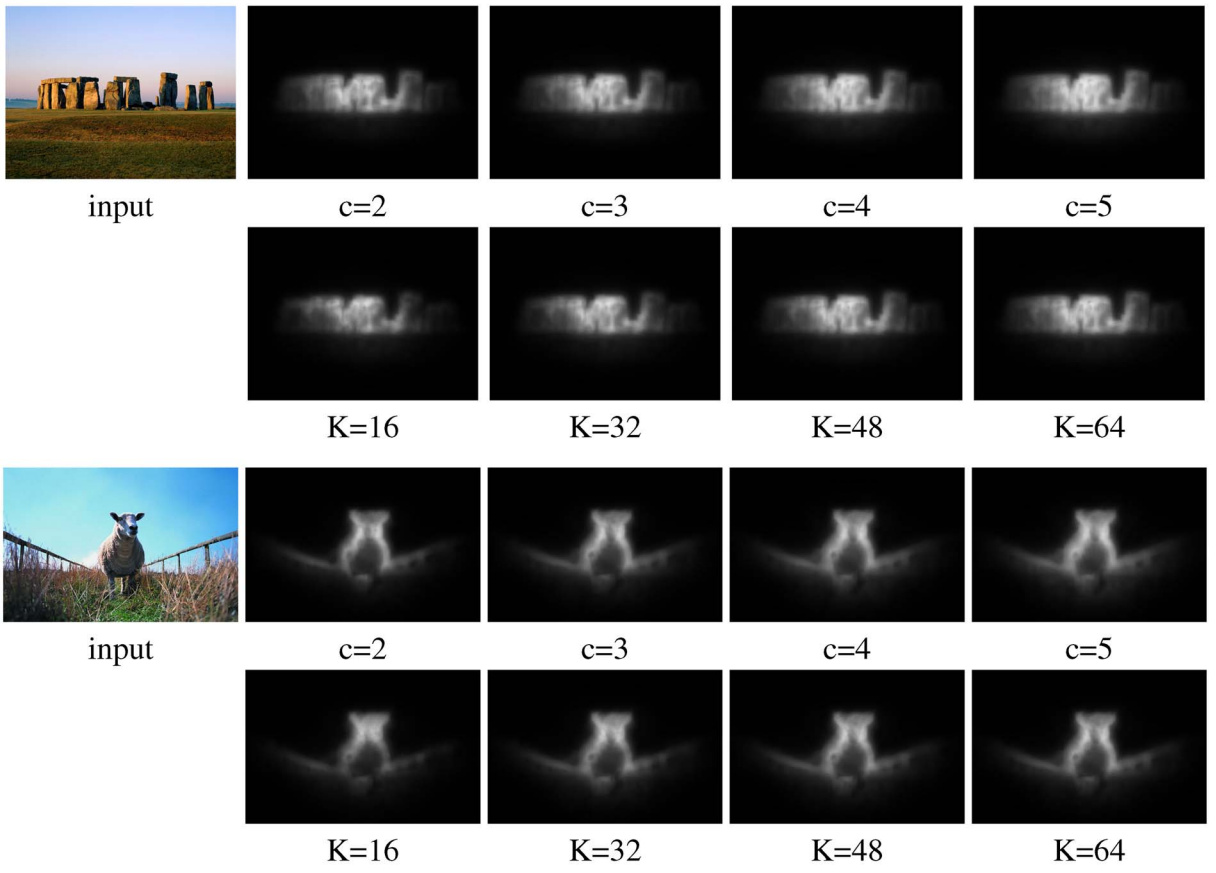


Fig. 13. Robustness to the parameters. Changing the parameters c and K has only a slight effect on the results.

varying c and K. It can be seen that our algorithm is very robust to the values of these parameters. While the saliency maps change slightly, the overall ROC curves end up almost identical, and therefore they are excluded. Fig. 14 shows the ROC curve for changing the number of scales R. The same results were obtained when using two or four scales. If only one scale is used, the results are inferior. This justifies our multiscale approach.

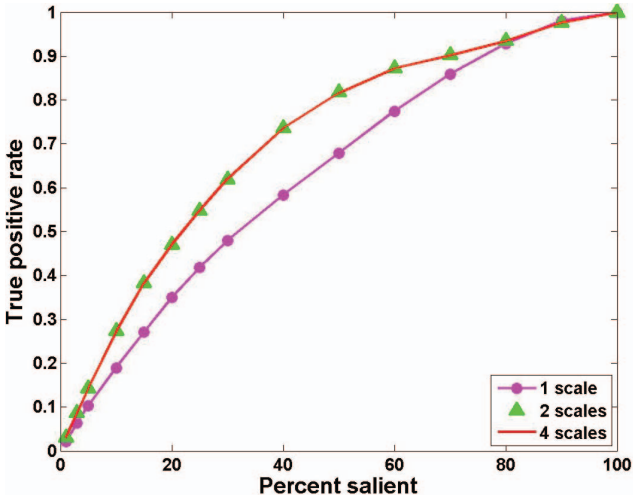


Fig. 14. The effect of the number of scales R on the database of [14]. A single scale produces inferior results.

6 GPU Implementation

Our saliency algorithm uses extensively the K-Nearest Neighbor (KNN) algorithm. Although simple in calculation, KNN has a very high computational complexity. Conceptually, for each point in $P = \{p_1, p_2, \ldots, p_{N_p}\}$, the distances to every point in $Q = \{q_1, q_2, \ldots, q_{N_q}\}$, $p_i, q_j \in R^d$, must be computed and then sorted to determine the K nearest neighbors. This represents a polynomial complexity in terms of point set size $(N_p \times N_q)$. To speed up the saliency computation one could replace the exact KNN with an approximate solution. Several approximate solutions to KNN have been proposed in order to reduce computation time [3], [25]. For instance, Arya et al. [3] partition the point sets using a kd-tree structure, and only compute distances within nearby volumes. The reduction in computation time comes at the cost of accepting errors in the returned nearest neighbors. In our case, unfortunately, even the more efficient implementations of approximate KNN are not able to run in real time (or near real time) even when run on the latest CPUs.

Recent opening of GPUs to general purpose computation (GPGPU) introduced a powerful platform with parallel calculation capabilities. Luckily, its highly parallel nature (the distances between pairs of points are independent) makes KNN a perfect candidate for being implemented on GPU. One such implementation was introduced in [7] and [8], using NVIDIA's C-based API Compute Unified Device Architecture (CUDA). In [7] and [8], an implementation of KNN search on GPU was proposed which achieves a speed

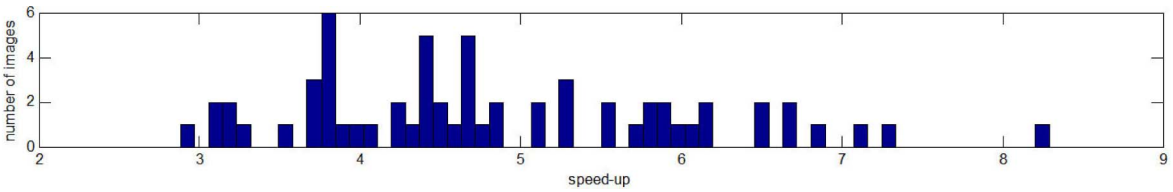


Fig. 15. GPU speedup.

increase by up to two orders of magnitude compared to CPU-based implementations. Garcia et al. [7], [8] have recently published online KNN CUDA source code [11]. We have incorporated their GPU code into our implementation resulting in a significant increase in speed.

Our algorithm is implemented in MATLAB on i7-860 CPU PC with 8 GB RAM and GeForce GTX 460 Graphics card. We compare between two versions of KNN, a CPU-based KNN (C++ ANN [2]) and GPU-based KNN [11]. We ran both versions on the database of [14] compiled of 62 images. The mean runtime of GPU implementation was 4.8 seconds per image, five times faster than the CPU implementation. Fig. 15 shows the speedup distribution obtained.

7 Application to Retargeting

Many applications require saliency maps as input. In this section, we show via the retargeting application that our proposed context-aware saliency is beneficial.

Image retargeting aims at resizing an image by expanding or shrinking the noninformative regions [4], [29], [24]. Therefore, retargeting algorithms rely on the availability of saliency maps which accurately detect all the salient image details.

Using context-aware saliency for retargeting could assure that the dominant objects, as well as their meaningful neighborhoods, will remain untouched in the resized image. Distortions, if and when introduced, will exist only in regions of lower significance.

Seam carving is a popular retargeting technique that repeatedly carves out seams in a certain direction [29]. To get pleasing results, removal/addition of seams should not introduce salient features. The selection and order of seams attempt to protect the content of the image, according to the saliency map. We ran the original code of [29] and compared their results with those produced after replacing their saliency map with ours.

Fig. 16 presents a couple of results. Differently from [29], our saliency guarantees that the salient objects (the fish and the men) are not distorted. The improved results can be explained by comparing the saliency maps. In the saliency maps of [29], the background appears important due to the edges it contains. Consequently, the seam carving algorithm prefers to carve out parts of the salient object. On the other hand, our saliency maps differentiate between the non-salient background and the salient object and its close salient background. Both are maintained after resizing, resulting in eye-pleasing images.

Further comparisons are provided in Fig. 17, where the saliency map of seam-carving is replaced by that of [21]. In [21], the dominant object is detected, but the object details

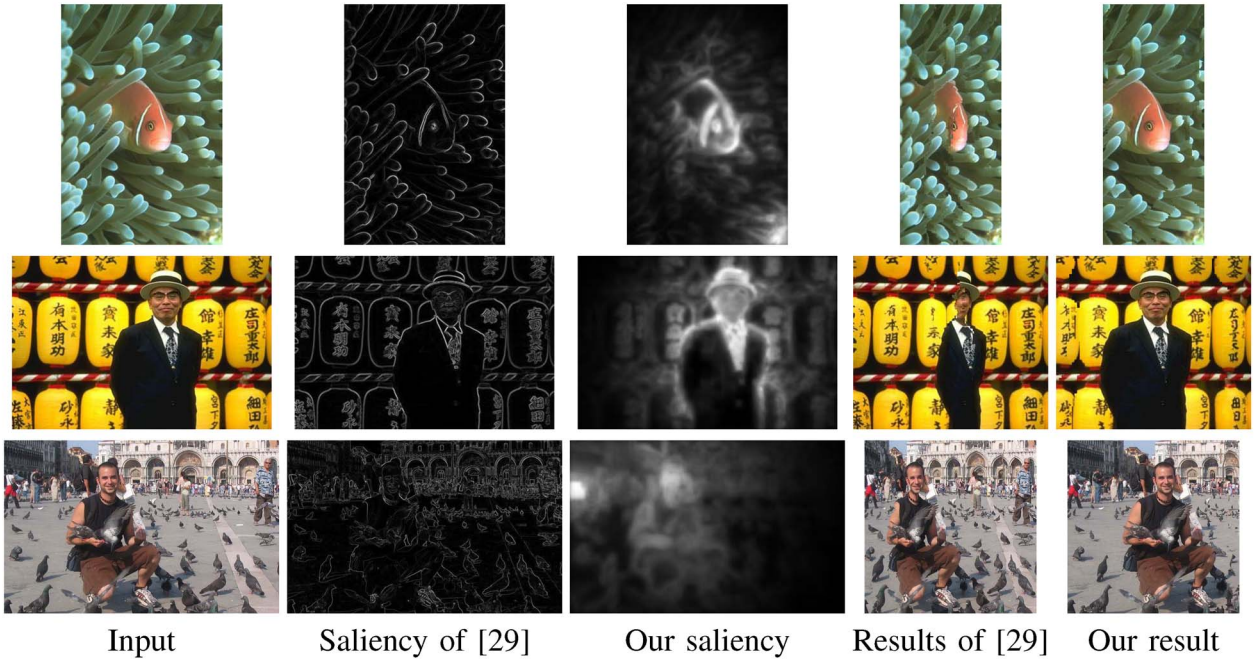


Fig. 16. Seam carving of 100 (top two rows) and 300 (bottom row) “vertical” lines. The salient objects are distorted by Rubinstein et al. [29], in contrast to our results.

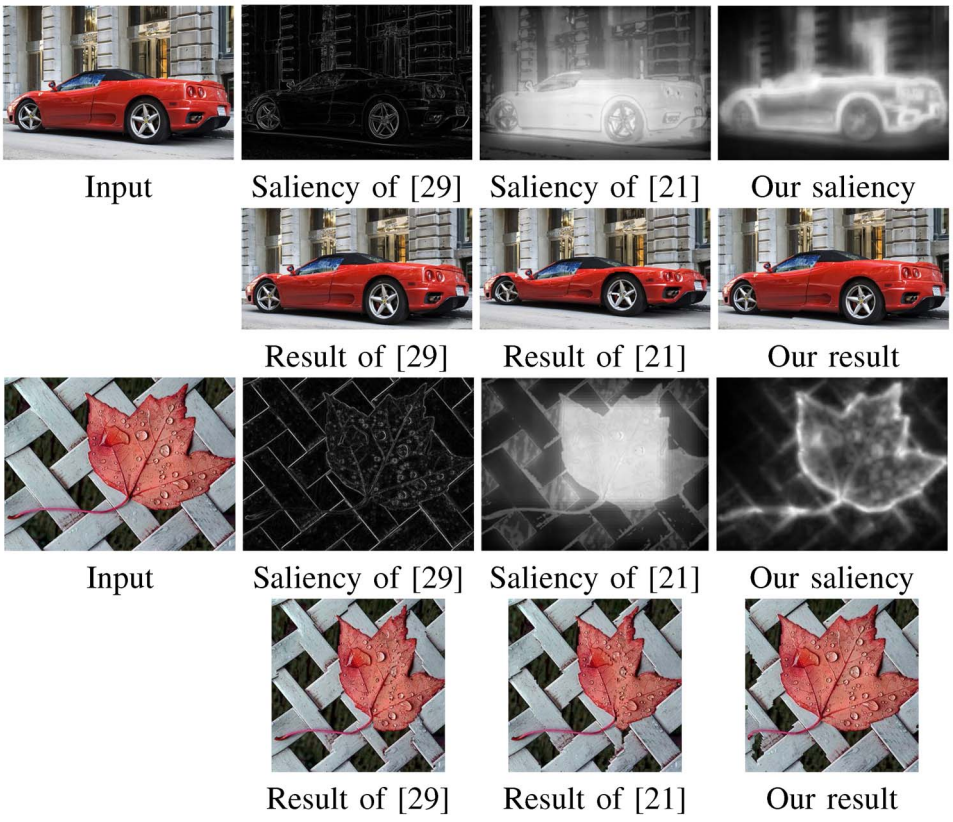


Fig. 17. In our saliency maps, the details of the car and the leaf are detected more accurately; hence they are not distorted by retargeting.

are not outlined accurately. Therefore, carving out seams through the car and the leaf (the dominant objects) does not generate new salient features and hence these seams are selected. In contrast, when our saliency is used, seams through the car and leaf would introduce salient features. In this case, these seams are avoided, leaving the objects untouched and resulting in less distortions.

8 Conclusion

This paper proposes a new type of saliency—context-aware saliency—which detects the important parts of the scene. This saliency is based on four principles observed in the psychological literature: local low-level considerations, global considerations, visual organizational rules, and high-level factors. The paper further presents an algorithm for computing this saliency.

There exists a variety of applications where the context of the dominant objects is just as essential as the objects themselves. This paper evaluated the contribution of context-aware saliency in two such applications—retargeting and summarization. In the future, we intend to learn the benefits of this saliency in more applications, such as image classification and thumbnailing.

Acknowledgments

This work was supported by the Fund for the Promotion of Research at the Technion and by the Ollendorff Foundation.

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-Tuned Salient Region Detection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1597-1604, 2009.

[2] ANN, http://www.cs.umd.edu/mount/ANN/, 2011.

[3] S. Arya, D. Mount, N. Netanyahu, R. Silverman, and A. Wu, "An Optimal Algorithm for Approximate Nearest Neighbor Searching Fixed Dimensions," *J. ACM,* vol. 45, no. 6, pp. 891-923, 1998.

[4] S. Avidan and A. Shamir, "Seam Carving for Content-Aware Image Resizing," *ACM Trans. Graphics,* vol. 26, no. 3, p. 10, 2007.

[5] O. Boiman and M. Irani, "Detecting Irregularities in Images and in Video," *Int'l J. Computer Vision,* vol. 74, no. 1, pp. 17-31, 2007.

[6] N. Bruce and J. Tsotsos, "Saliency Based on Information Maximization," *Advances in Neural Information Processing Systems,* vol. 18, pp. 155-162, 2006.

[7] V. Garcia, "Suivi Dobjets Dintrt dans une Squence Dimages: Des Points Saillants aux Mesures Statistiques," PhD thesis, Universit de Nice, 2008.

[8] V. Garcia, E. Debreuve, and M. Barlaud, "Fast k Nearest Neighbor Search Using GPU," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.

[9] S. Goferman, A. Tal, and L. Zelnik-Manor, "Puzzle-Like Collage," *Computer Graphics Forum,* vol. 29, pp. 459-468, 2010.

[10] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-Aware Saliency Detection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 2376-2383, 2010.

[11] GPU, http://www.i3s.unice.fr/~creative/KNN/, 2011.

[12] C. Guo, Q. Ma, and L. Zhang, "Spatio-Temporal Saliency Detection Using Phase Spectrum of Quaternion Fourier Transform," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1-8, 2008.

[13] J. Harel, C. Koch, and P. Perona, "Graph-Based Visual Saliency," *Advances in Neural Information Processing Systems,* vol. 19, pp. 545-552, 2007.

[14] X. Hou and L. Zhang, "Saliency Detection: A Spectral Residual Approach," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1-8, 2007.

[15] L. Itti and C. Koch, "Computational Modelling of Visual Attention," *Nature Rev. Neuroscience,* vol. 2, no. 3, pp. 194-204, 2001.
[16] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 11, pp. 1254-1259, Nov. 1998.
[17] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to Predict Where Humans Look," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 2106-2113, 2009.
[18] C. Koch and T. Poggio, "Predicting the Visual World: Silence Is Golden," *Nature Neuroscience,* vol. 2, pp. 9-10, 1999.
[19] K. Koffka, *Principles of Gestalt Psychology.* Routledge & Kegan Paul, 1955.
[20] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A Coherent Computational Approach to Model Bottom-up Visual Attention," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 5, pp. 802-817, May 2006.
[21] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Shum, "Learning to Detect a Salient Object," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2007.
[22] Y. Ma and H. Zhang, "Contrast-Based Image Attention Analysis by Using Fuzzy Growing," *Proc. 11th ACM Int'l Conf. Multimedia,* pp. 374-381, 2003.
[23] E. Nowak, F. Jurie, and B. Triggs, "Sampling Strategies for Bag-of-Features Image Classification," *Proc. Ninth European Conf. Computer Vision,* pp. 490-503, 2006.
[24] Y. Pritch, E. Kav-Venaki, and S. Peleg, "Shift Map Image Editing," *Proc. 12th IEEE Int'l Conf. Computer Vision,* pp. 151-158, 2009.
[25] Z.W.M.C.Q. Lv, W. Josephson, and K. Li, "Multiprobe LSH: Efficient Indexing for High-Dimensional Similarity Search," *Proc. 33rd Int'l Conf. Very Large Data Bases,* pp. 950-961, 2007.
[26] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting Salient Objects from Images and Videos," *Proc. 11th European Conf. Computer Vision,* pp. 366-379, 2010.
[27] C. Rother, L. Bordeaux, Y. Hamadi, and A. Blake, "Autocollage," *ACM Trans. Graphics,* vol. 25, no. 3, pp. 847-852, 2006.
[28] C. Rother, V. Kolmogorov, and A. Blake, "'GrabCut': Interactive Foreground Extraction Using Iterated Graph Cuts," *ACM Trans. Graphics,* vol. 23, no. 3, pp. 309-314, 2004.
[29] M. Rubinstein, A. Shamir, and S. Avidan, "Improved Seam Carving for Video Retargeting," *ACM Trans. Graphics,* vol. 27, no. 3, p. 16, 2008.
[30] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is Bottom-Up Attention Useful for Object Recognition?," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 2, 2004.
[31] H. Seo and P. Milanfar, "Static and Space-Time Visual Saliency Detection by Self-Resemblance," *J. Vision,* vol. 9, no. 12, pp. 1-27, 2009.
[32] B. Suh, H. Ling, B.B. Bederson, and D.W. Jacobs, "Automatic Thumbnail Cropping and Its Effectiveness," *Proc. 16th Ann. ACM Symp. User Interface Software and Technology,* pp. 95-104, 2003.
[33] A. Treisman and G. Gelade, "A Feature-Integration Theory of Attention," *Cognitive Psychology,* vol. 12, no. 1, pp. 97-136, 1980.
[34] P. Viola and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," *Proc. IEEE Computer Vision and Pattern Recognition,* 2001.
[35] D. Walther and C. Koch, "Modeling Attention to Salient Proto-Objects," *Neural Networks,* vol. 19, no. 9, pp. 1395-1407, 2006.
[36] J. Wolfe, "Guided Search 2.0: A Revised Model of Visual Search," *Psychonomic Bull. and Rev.,* vol. 1, no. 2, pp. 202-238, 1994.

Stas Goferman received the BSc degree (summa cum laude) in electrical engineering from the Technion in 2001. After graduating, he served for 10 years with the Technological Department of the Israeli Ministry of Defense. In parallel to his service, he received the MSc degree (cum laude) in electrical engineering from the Technion in 2009. He held various R&D positions, starting from hardware engineer and then a systems engineer of a multi-disciplinary project which included development of complex hardware and software elements. Later on, he served as a senior engineer in the electronics reliability and quality division. At his last position he held an entrepreneurship position, successfully leading innovative projects. His research includes analysis of visual data and image summarization techniques. In 2004, 2005, and 2007, he received National Technological Excellence Awards. In 2010, he received a National Innovation Award.

Lihi Zelnik-Manor received the BSc degree in mechanical engineering from the Technion in 1995, where she graduated summa cum laude, and the MSc (with honors) and PhD degrees in computer science from the Weizmann Institute of Science in 1998 and 2004, respectively. After graduating, she worked as a postdoctoral fellow in the Department of Engineering and Applied Science at the California Institute of Technology (Caltech). Since 2007 she has been a senior lecturer in the Electrical Engineering Department at the Technion. Her research focuses on the analysis of dynamic visual data, including video analysis and visualizations of multiview data. Her awards and honors include the Israeli high-education planning and budgeting committee (Vatat) three-year scholarship for outstanding PhD students, and the Sloan-Swartz postdoctoral fellowship. She also received the best Student Paper Award at the IEEE Shape Modeling International Conference 2005 and the AIM@SHAPE Best Paper Award 2005. She is a member of the IEEE.

Ayellet Tal received the BSc degree (summa cum laude) in mathematics and computer science and the MSc degree (summa cum laude) in computer science, both from Tel-Aviv University, and the PhD degree in computer science from Princeton University. Currently, she is working as an associate professor in the Department of Electrical Engineering at the Technion and is the founder of the Laboratory of Computer Graphics and Multimedia. Her research interests include computer graphics, information and scientific visualization, computational geometry, and multimedia. She served as the program chair of the ACM Symposium on Virtual Reality, Software, and Technology (VRST) and as the chair of Shape Modeling International (SMI). She has also served on the program committees of all the leading conferences in computer graphics. She is an associate editor of *Computers and Graphics* and on the editorial board of the journal *Computer Graphics Forum* (*CGF*). She has also edited several special issues of various journals. She is a recipient of the Henry Taub Prize for Academic Excellence, the Google Research Award, as well as several grants from ISF, MOST, the sixth European R&D Program, and others.