Image Saliency: From Intrinsic to Extrinsic Context

Meng Wang, Janusz Konrad, Prakash Ishwar
Dept. of Electrical and Computer Eng., Boston University
Boston, MA 02215
{wangmeng,jkonrad,pi}@bu.edu

Kevin Jing, Henry Rowley
Google Research.
Mountain View, CA 94043
{jing,har}@google.edu

Abstract

We propose a novel framework for automatic saliency estimation in natural images. We consider saliency to be an anomaly with respect to a given context that can be global or local. In the case of global context, we estimate saliency in the whole image relative to a large dictionary of images. Unlike in some prior methods, this dictionary is not annotated, i.e., saliency is assumed unknown. In the case of local context, we partition the image into patches and estimate saliency in each patch relative to a large dictionary of un-annotated patches from the rest of the image. We propose a unified framework that applies to both cases in three steps. First, given an input (image or patch) we extract k nearest neighbors from the dictionary. Then, we geometrically warp each neighbor to match the input. Finally, we derive the saliency map from the mean absolute error between the input and all its warped neighbors. This algorithm is not only easy to implement but also outperforms state-of-the-art methods.

1. Introduction

If you show a group of people the image in Fig. 1(a) and ask them to annotate parts of it that they consider salient (scale 0 to 1), you will likely get a distribution of saliency scores that is roughly consistent with the images in Fig. 1(b)–(d). If, however, you first show them the images in Fig. 1(e) and then ask them to annotate parts of Fig. 1(a) that they consider salient *in relation to* the images of Fig. 1(e), very likely the "missing leg" will stand out as the most salient part, as shown in Fig. 1(f). We have developed a simple unified framework for image saliency estimation that covers both of these scenarios and more.

Our thesis is that saliency is meaningful only in relation to a context and work to-date has implicitly aimed to capture what may be termed as intrinsic saliency, i.e., saliency in the context of local image structure (even when based on a large dictionary of images). We hold that regions of an image that should be considered salient in relation to a dictionary of images or local image patches are not those

Figure 1. For an input image (a), traditional saliency estimation algorithms, such as those by Itti98 [14] (b) and Hou08 [10] (c), only aim to capture what may be termed as intrinsic saliency. In contrast, the proposed algorithm can produce not only an intrinsic saliency map (d) in the local context of a dictionary of image patches but also an extrinsic saliency map (f) in the global context of a dictionary of images containing only two-legged humans (e).

which occur in great abundance within the dictionary but precisely those which are unusual. Thus a salient region is a region which is anomalous relative to a dictionary. We are, in essence, advocating a definition of saliency which not only encompasses the implicit traditional notion as a special case, namely when the context is local image structure, but which significantly expands the scope and utility of the traditional notion as illustrated by the saliency of the missing leg (Fig. 1(f)) of the one-legged man (Fig. 1(a)) in a universe of two-legged men (Fig. 1(e)).

Our proposed saliency estimation algorithm consists of three steps as illustrated in Fig. 2. In the global (extrinsic) case, we estimate saliency in the whole image: the input image is evaluated for saliency against a dictionary of images, that has *not* been annotated for presence or location of saliency, unlike in [20]. In the local (intrinsic) case,

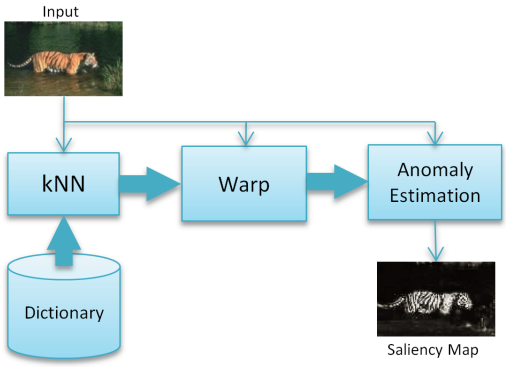


Figure 2. Block diagram of the proposed (extrinsic) saliency estimation algorithm based on 3 steps: (1) kNN search to retrieve k dictionary elements similar to the input, (2) geometric warping to align neighbors to the input, (3) anomaly estimation.

we estimate saliency on a patch-by-patch basis: each input patch in an image is evaluated for saliency against the dictionary of remaining patches from the same image or parts thereof. Depending on the size and composition of the dictionary, immediate-neighborhood saliency or full-image saliency can be estimated.

In both local and global cases, first a kNN search *vis-à-vis* the input (patch or image) is executed on the dictionary. The extracted k neighbors have a visual appearance close to that of the input but size, shape and position of objects may significantly differ among them. We address this in the second step by geometrically warping each neighbor to match the input. The warped neighbors are likely to exhibit similar background (luminance, color, location) to that in the input since the kNN search is based on global image properties; as the background is usually dominant in an image in terms of size, it provides a strong bias in the kNN search. However, most foreground areas are small and likely to vary among neighbors if the context dictionary is sufficiently rich. Most of them will also differ from foreground areas in the input. We leverage this in the third step by computing the mean absolute error between the warped neighbors and the input. The saliency map of the input is then assumed to be a normalized version of this error. Due to the similarity of the background between the input and the warped neighbors and, at the same time, likely variability of the foreground among the warped neighbors, the saliency is naturally collocated with large mean errors. In other words, it is anomalous with respect to the foregrounds in the warped neighbors.

We would like to point out that in one specific instantiation of our framework, we use a dictionary of 100 million online images to infer the global context. We solve the obvious issue of kNN search complexity by leveraging massive parallelism.

Despite its simplicity, our framework performs very well. Both the local- and global-context variants outperform most of the state-of-the-art saliency estimation methods on Berkeley [21] and MSRA [19] databases. Furthermore, a combined local/global variant performs even better outperforming all of the methods we have tested.

In summary, the contributions of this paper are twofold:

1. We develop a novel algorithmic framework for saliency estimation based on kNN search and warping by using either local context (intrinsic to the image) or global context (extrinsic to the image), or a combination of the two.

2. In the global context, we leverage the wealth of unannotated image and video data available on-line, for accurate saliency estimation.

The rest of the paper is organized as follows. In Section 2, we review related work on saliency estimation. In Section 3, we describe our method in detail. Experimental results and comparisons with other methods are presented in Section 4. We conclude with a discussion in Section 5.

2. Related work

The process of visual attention has been extensively studied in psychology, neural systems, and computer vision (e.g., [15, 28]). At the highest level, two mechanisms are usually considered when explaining visual attention in humans: *bottom-up*, driven by saliency, and *top-down*, volitionally-controlled attention bias [22]. Current computational models of visual attention are primarily based on the bottom-up mechanism, i.e., saliency maps, and have been inspired by the work of Koch and Ullman [15].

Most approaches to the modeling and localization of visual attention to date [1, 12, 13, 14, 17, 20, 29] are based on a feature integration framework [27]. Typically, these approaches consist of three stages. First, various low-level visual features, such as luminance, color, edge orientations and texture patterns are extracted from the image, often at multiple scales. Then, "activation maps" are computed from the extracted visual features, for example, by means of a center-surround operation that emulates the visual receptive fields [14], Shannon's self-information measure that is inspired by the primate visual cortex [5], or biologically plausible graph-based random walks [9]. Finally, normalization and fusion (linear or non-linear) of multiple "activation maps" is performed to yield a saliency map.

However, the semantic content of a scene, co-occurrences of objects and task constraints have been shown to play a key role in visual attention bias [2, 6, 4, 16]. Observers can be implicitly cued to a target location by global properties of the image, like background color or texture [4, 16, 23]. Different from feature integration model, contextual information can also be used for saliency estima-

tion, for example by exploiting the relationship between co-occurring objects in real-world environment [22]. Clearly, context plays a pivotal role in human visual attention.

The performance of bottom-up methods is strongly dependent on the specific choice of features. In an attempt to address this dependence, Liu *et al.* [19] have proposed to solve the feature selection problem by means of training on a large accurately-labeled image database. Clearly, this use of context is supervised and thus not scalable.

Oliva *et al.* [26] have proposed a computational model of attention guidance that combines bottom-up saliency with top-down bias provided by viewers in a supervisory manner. They have shown that context can provide a low-complexity object detection by pre-selecting relevant image regions. However, their experimental results were limited to a very specific class of objects, e.g., pedestrians.

Context has been recently emphasized in saliency detection by Goferman *et al.* [7]. The proposed model combines immediate and distant contexts (neighboring versus far-away patches) by weighing a color difference metric with the inverse of a distance metric. Consequently, a patch with distinct color, as compared to its neighbors, is more salient than it is against distant patches (with the same colors as the neighbors). Note that in this case the comparisons are performed only within the image itself.

A method proposed by Boiman and Irani [3], like our method, treats saliency as anomaly with respect to a dictionary. However, while they use from the same image as the dictionary, we use patches for local context and a dictionary of complete images for global context. Also, we introduce geometric warping to align k most similar patches/images from the dictionary with the input patches/images. Most importantly, however, Boiman and Irani's method cannot be easily extended to large unstructured dictionaries such as the Google image database since in their method a patch will be labeled as not salient (normal) if there exists even one similar patch in the dictionary; once the dictionary is sufficiently rich, as in our case, the method will not label any patch as salient.

Perhaps closest to our work is the recent approach proposed by Marchesotti *et al.* [20]. Similarly to our approach, the context is retrieved from a database of images using kNN search. However, unlike in our approach, this database is assumed manually annotated for saliency (bounding boxes). The saliency of the k nearest neighbors is transferred to the query in form of a bounding box. Clearly, the method is semi-supervised and, furthermore, does not produce detailed saliency maps.

Our work has been largely inspired by recent data-driven approaches in computer vision [24, 25, 30, 31]. It has been shown that given a large data set, there always exist visually similar images to a query. Such neighbors have been shown to be useful in object recognition tasks. In the case of our approach, the neighbors provide context, i.e., relationship, between objects or between objects and background.

Essential to our framework is the warping of nearest neighbors in order to better match the query. To accomplish this, we borrow from the work of Liu *et al.* [18] that proposes a robust optical flow algorithm to align similar objects in different images. We use this method to generate context (mostly background) matched to the query.

3. Saliency estimation algorithm

As explained in the introduction, we view saliency as anomaly relative to a context. The context could be local or global. A context is instantiated by a dictionary of image patches. In the local case, image patches are local neighborhoods of the input image itself. Since the dictionary is composed of regions internal to the input image, we call this **intrinsic saliency estimation**. In the global case, image patches are entire images that share a global context with the input image and are therefore relevant for estimating saliency. Since we use information from outside the input image, we call this **extrinsic saliency estimation**. Both intrinsic and extrinsic saliency estimation share a common computational framework that we motivate and describe in what follows. We will first focus on the extrinsic case and then describe how it can be adapted in a straightforward manner to the intrinsic case at the end of this section.

3.1. Extrinsic saliency estimation

The high-level block diagram of our overall algorithm for extrinsic saliency estimation is depicted in Fig. 2. A more detailed view is presented in Fig. 3. Our algorithm for saliency estimation is motivated by anomaly detection in video analytics where an incident is picked out as anomaly if it cannot be "explained" by previously seen incidents. Our algorithm consists of three steps: 1) using a k-nearest-neighbors (kNN) algorithm to discard images from the dictionary that are irrelevant for saliency estimation, 2) using a warping algorithm to geometrically align the k nearest neighbors to the input, and 3) estimating saliency as the average absolute error, suitably normalized to the range $[0, 1]$, between the input and the warped nearest neighbors. The underlying intuition and details of these three steps are discussed below.

3.1.1 KNN image retrieval

The images in a large dictionary can be partitioned into two groups: those which are relevant for determining saliency in a given input image and those that are irrelevant. Images that are totally dissimilar with respect to the input image are not very useful for estimating saliency because relative to them, almost all points in the input image would roughly be "equally salient". In other words, these images would only

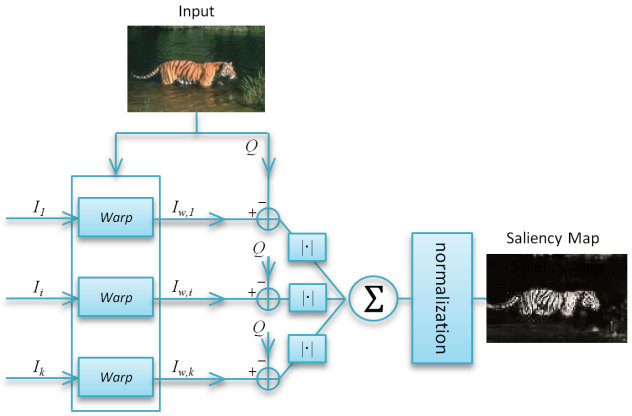


Figure 3. Detailed view of the image warping and anomaly estimation steps for extrinsic (global) saliency estimation from Fig. 2. The k nearest neighbor images $I_i, i = 1, \ldots, k$, of the input image Q are respectively warped, using the SIFT flow algorithm [18], to the images $I_{w,i}, i = 1, \ldots, k$ which are approximately "aligned" with Q. The absolute error images $|I_{w,i} - Q|$ are averaged and normalized to generate the final saliency map.

contribute global "noise-terms" which are uninformative for estimating which regions in the input image have higher or lower saliency (on a $[0, 1]$ scale) compared to other parts of the input image. One method for selecting only a small useful subset of saliency-relevant images from a large dictionary is to select only the k images that are closest to the input where closeness is measured by some distance function which captures global image properties such as color, texture, edges, etc. The selection of a smaller subset of images for saliency estimation provides the added practical benefit of computational tractability when the size of the dictionary is very large.

We use the distance function from [8] which is based on a weighted Hamming distance between binary hashes of features that capture global image structure (we refer the reader to [8] for the details). The binary hash representation of images and the associated weighted Hamming distance function have proved to be concise enough for fast search while retaining enough discriminative power to reliably distinguish between different scenes.

To ensure that our image dictionary is sufficiently large and diverse for retrieving globally visually similar images for a wide spectrum of queries, we built it out of 100 million natural images from online resources. For a dictionary this large, exact kNN search is infeasible. We use a massively parallelized implementation of the approximate kNN algorithm developed in [25] which completes the entire search and retrieval under one second.

Fig. 4 shows search results for a particular input image. An examination of the search results shows that the neighboring images indeed are within the same global context as the input image.

Figure 4. The 42 nearest neighbors (right) of the input image (left) retrieved by the algorithm of [25] using the image distance function of [8]. The k nearest neighbors of an input image in a large dictionary correspond to images that are most relevant to saliency estimation.

Like all kNN based methods, we need to choose a value of k. Typically, the average distance between an input and its k-th nearest neighbor decays with increasing values of k. A sharp peak around 0 indicates that only a few images are very close to the input. The variation of saliency estimation performance for different values of k is discussed in Section 4. We use the nominal value of $k = 20$ in all our experiments for extrinsic saliency estimation.

3.1.2 Image warping

Although the neighboring images contain the same global context, the objects they contain may not be geometrically well aligned with those in the input image. For instance, it may happen that both the input image and its neighbors have similar horizons (in an outdoor scene) and similar-looking objects but their positions and orientations may be slightly different. Such misalignments, if unaccounted for, may lead to poor estimates of saliency. We address this issue by warping the neighbors so that they are well aligned with the input image before estimating saliency. We use the SIFT flow algorithm of [18] to geometrically warp each neighbor image $I_i, i = 1, \ldots, k$, retrieved by the kNN search, to align it to the input image as illustrated in Fig. 5. The resulting warped image is denoted in Fig. 3 as $I_{w,i}$.

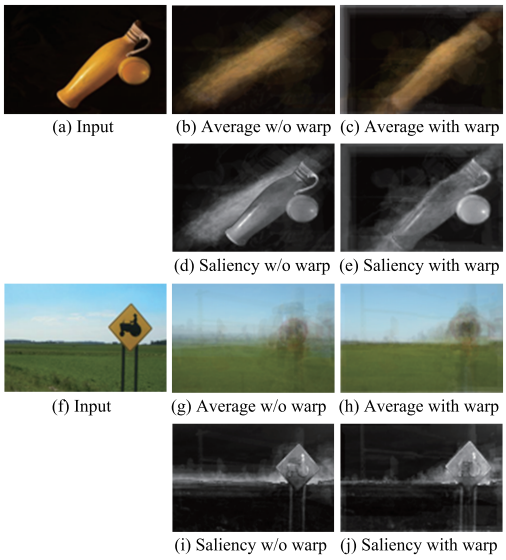
Fig. 6 illustrates how the image warping affects the saliency estimation. Without warping, the simple average of all k nearest neighbors is both more blurred and misaligned relative to the image than with warping.

3.1.3 Anomaly estimation

Each warped neighbor provides an approximation to the input image. Some regions of the input image, such as the background, may be well approximated by many warped neighbors from the dictionary, while other regions may not be well approximated by any. Regions that are better approximated by many warped neighbors are less salient as

(a) (b) (c)

Figure 5. Illustration of image warping *via* the SIFT flow algorithm [18]. (a) Input image; (b) 3 neighboring images retrieved by the kNN method; (c) Neighboring images after warping using SIFT flow.



(a) Input (b) Average w/o warp (c) Average with warp

(d) Saliency w/o warp (e) Saliency with warp

(f) Input (g) Average w/o warp (h) Average with warp

(i) Saliency w/o warp (j) Saliency with warp

Figure 6. Illustration of the importance of image warping. Left: input images; middle: averaged neighbor images and corresponding saliency maps without warping; right: averaged neighbor images and corresponding saliency maps with warping. Notice the blurring and misalignment of the vase and horizon without warping and the corresponding misestimation of saliency at those locations.

they are commonplace among the relevant images in the dictionary. Regions where the approximation is poor are more salient. We measure the saliency $S(x,y)$ at point (x,y) of the input image $Q(x,y)$ as the absolute error $|I_{w,i}(x,y) - Q(x,y)|$ averaged across all k warped neighbors $I_{w,i}(x,y), i = 1,\ldots,k$ and normalized to the range $[0,1]$. Mathematically,

$$\begin{aligned} E(x,y) &:= \frac{1}{k}\sum_{i=1}^{k} |I_{w,i}(x,y) - Q(x,y)| \\ S(x,y) &:= E(x,y)/\max_{x,y} E(x,y) \end{aligned}$$

Both the size and diversity of the dictionary contribute to the accuracy of saliency. The size of a dictionary is related to its redundancy and its diversity is related to its rank or intrinsic dimensionality. At one extreme, if the dictionary is extremely diverse but relatively small so that any two elements are totally dissimilar, then any input image will be well-approximated by only a few saliency-relevant warped neighbors and consequently almost all parts of the input image would be estimated as salient. At the other extreme, if the dictionary is very large but not diverse so that all elements look alike, then if the input is similar to any element then no part of it would be salient whereas if it is dissimilar to any element then almost all parts of the input image would be estimated as salient. A sufficiently large and diverse dictionary is therefore ideal. This is akin to the bias-variance tradeoff in regression problems. Instead of using the mean absolute error, one could use the mean squared error or other measure of goodness of fit. We have found that they produce roughly similar results.

3.2. Intrinsic saliency estimation

The extrinsic saliency estimation algorithm described previously is easily adapted to the scenario in which the only image available for saliency estimation is the input image, i.e., there is no external image dictionary available. The idea is to use the set of all local image patches as a dictionary and estimate saliency as before for each local patch of the input image relative to the dictionary. In our experiments, image patches are 16×16 non-overlapping blocks.[1] Then, for each block that we test for saliency, we use *all the blocks*, i.e., k = dictionary-size, warp them and calculate the difference and determine the saliency as we did in Sections 3.1.2 and 3.1.3. The saliency map of the input image is then given by the composition of the saliency maps of the non-overlapping blocks.

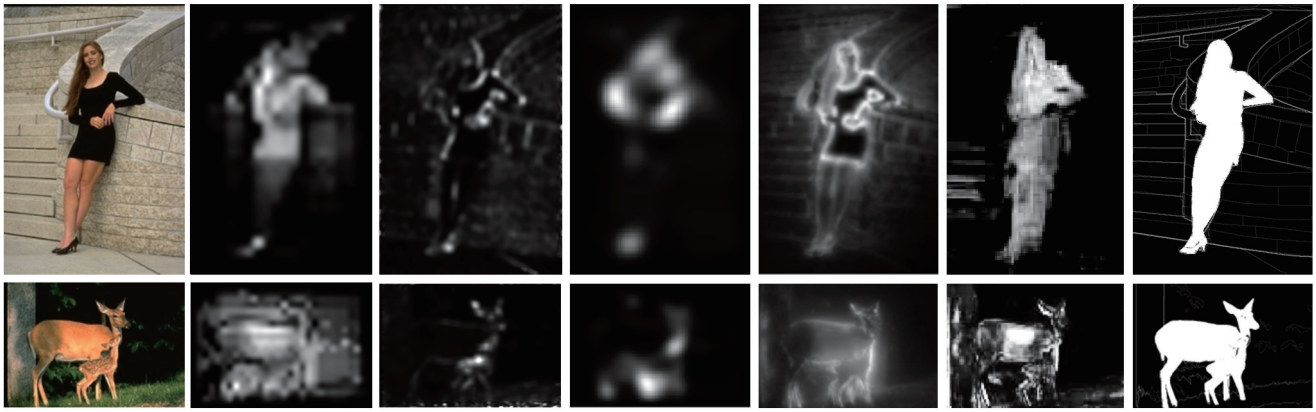
4. Results

We have tested the proposed methods and compared them with state-of-the-art saliency estimation algorithms on two data sets: Berkeley segmentation database (BSDB) [21] and MSRA salient object database [19].

The Berkeley segmentation database contains 300 natural images with object boundaries traced manually by a number of viewers. Depending on the image, individual boundaries may mostly overlap or be largely disjoint. We selected 50 images with clearly delineated foreground objects and we filled in the image area surrounded by the boundaries to obtain a binary ground truth with the foreground labeled as 1 and the rest labeled as 0. Some of the images used and the derived ground truths are shown in Figs. 7 and 9 in the first and last columns, respectively.

The MSRA salient object database consists of 2 sets of images: set A with 20,000 and set B with 5,000 natural

[1] One can also use overlapping blocks but this increases the computational complexity.



(a) Input (b) Itti98 (c) Hou07 (d) Hou08 (e) Goferman *et al.* (f) Our extrinsic (g) Ground truth

Figure 7. Comparison of saliency estimation results for various methods with our extrinsic approach for ($k = 20$) on the BSDB data set.

images. For each image, ground truth is known in form of a bounding box that reflects what typical viewers deem interesting in each image. We used the set B that contains images with less ambiguity as to the location of saliency. We filled in the bounding box with label 1 and the rest of the image with label 0 to create our binary ground truth.

In our extrinsic approach, we used 100 million online images as the dictionary with k, the number of nearest neighbors used, as high as 90, whereas in the intrinsic approach we used 16×16 blocks as image patches and k equal to the number of the remaining patches in the image.

In various tests we have compared our approach with methods proposed by Itti *et al.* [14], Hou and Zhang [11, 10], Achanta *et al.* [1] and the context-aware approach of Goferman *et al.* [7]. For the Itti *et al.*'s method we used source code from `http://www.saliencytoolbox.net`, and for Hou and Zhang's methods, Achanta *et al.*'s method and Goferman *et al.*'s method we used source code from the authors' web site respectively.

Fig. 7 compares these four methods with our extrinsic approach and the ground truth. Subjectively, our method performs at least as well as other methods: Itti *et al.*'s method lacks precise localization (large portions of the background are deemed to be likely salient as well), Hou and Zhang's 2007 method concentrates on object boundaries while picking up saliency in the background too, their 2008 method misses large parts of salient objects and produces diffuse maps, whereas the method of Goferman *et al.*, though quite precise, captures a lot of background detail too.

In an objective test, we have thresholded each continuous-valued saliency map using various thresholds between 0 and 1 to produce a binary saliency map (1 = above threshold, 0 = below threshold). Then, we computed an average of true positives and false positives

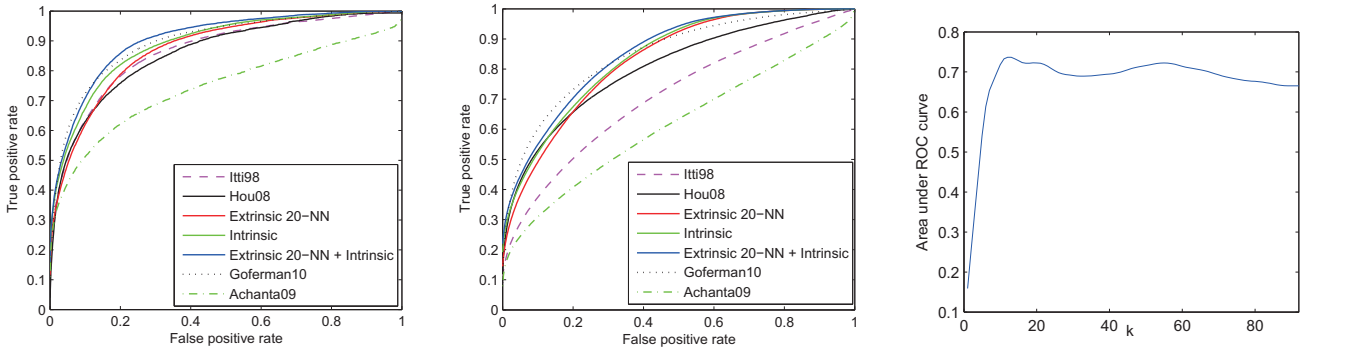
Method	BSDB	MSRA
Achanta09	0.7442	0.6743
Itti98	0.8641	0.6967
Hou08	0.8579	0.7934
Goferman10	0.8957	0.8437
Extrinsic 20-NN	0.8728	0.8289
Intrinsic	0.8881	0.8389
Extrinsic 20-NN + Intrinsic	0.9042	0.8515

Table 1. Area under the curves from Fig. 8(a-b).

vis-à-vis ground truth derived from the BSDB database (object shape) or MSRA database (rectangle). By varying the threshold, we produced true-positives-versus-false-positives curves shown in Fig. 8(a-b). Clearly, both intrinsic and extrinsic methods that we proposed perform as well or better than the methods of Achanta *et al.*[2], Itti *et al.* and Hou and Zhang 2008. Interestingly, our intrinsic method slightly outperforms the extrinsic method on these databases suggesting that local context may be often enough. However, a combined extrinsic/intrinsic method, that simply multiplies the saliency maps produced by the two methods, performs even better. In fact, it outperforms the recent method by Goferman *et al.* This may suggest that taking local *and* global context simultaneously is beneficial. Table 1 shows the area under each curve from Fig. 8(a-b).

The extrinsic results above have been obtained for kNN search with $k = 20$. For comparison, Fig. 8(c) shows average performance for the extrinsic method (area under the red curve in Fig. 8(a)) as a function of k. Clearly, the best performance occurs around $k = 15$, then drops and increases again around $k = 55$. There is no benefit from using large values of k. This happens because for large k's more dissimilar neighbors are allowed and as such they cannot accurately model the context.

[2]We do not use any post processing described in Achanta *et al.* [1], which perhaps explains the sub-par performance.



(a) BSDB (b) MSRA-B (c) Area under curve for extrinsic (BSDB)

Figure 8. Comparison of average performance of various methods on: (a) BSDB, and (b) MSRA-B data sets, and (c) area under the curve for our extrinsic method on BSDB data set (red curve in part (a)) as a function of k.

More experimental results for various methods are shown in Fig. 9. On most images, the methods by Itti *et al*. and Hou & Zhang produce saliency maps that concentrate only on parts of foregrounds, while our extrinsic method detects, with few exceptions, complete foregrounds. Although our saliency maps are quite close to the ground truth on many images, in some cases they are less accurate. However, note that the saliency maps shown are continuous in value (0 to 1) from which a binary mask or bounding box needs to be derived. Clearly, in order to accomplish this a segmentation or classification step is needed. If such a step exploits regularization *via* some prior models, the final binary map is likely to be very close to ground truth in many cases shown in Fig. 9.

5. Discussion and conclusions

The framework for saliency estimation that we proposed is conceptually simple and straightforward to implement, especially in the intrinsic case – only warping and averaging are needed. It is also computationally efficient, again especially in the intrinsic case with disjoint patches; the complexity, and performance, increase when patch overlap is allowed. Although intrinsic context is often enough, its combination with extrinsic context gives even better performance, though at the cost of increased computational complexity. Cumbersome today, a search through millions of on-line images for global context is likely to be a commodity task in the future on account of distributed computing in the cloud.

Acknowledgments

This work was partially supported by the National Science Foundation under award CNS-0721884. Part of this work was supported by and performed at Google Inc. (first author). The authors would like to thank Ming Zhao and other members of the Google Video team for help with this work.

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Ssstrunk. Frequency-tuned Salient Region Detection. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 418, 422

[2] M. Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617–629, August 2004. 418

[3] O. Boiman and M. Irani. Detecting irregularities in images and in video. *Int. J. Comput. Vision*, 74:17–31, August 2007. 419

[4] J. R. Brockmole and J. M. Henderson. Using real world scenes as contextual cues for search. *Visual Cognition*, 13:99–108, 2006. 418

[5] N. D. Bruce and J. K. Tsotsos. Saliency based on information maximization. In *Proc. Conf. Neural Inf. Proc. Systems*, 2005. 418

[6] M. M. Chun and Y. Jiang. Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, 36:28–71, July 1998. 418

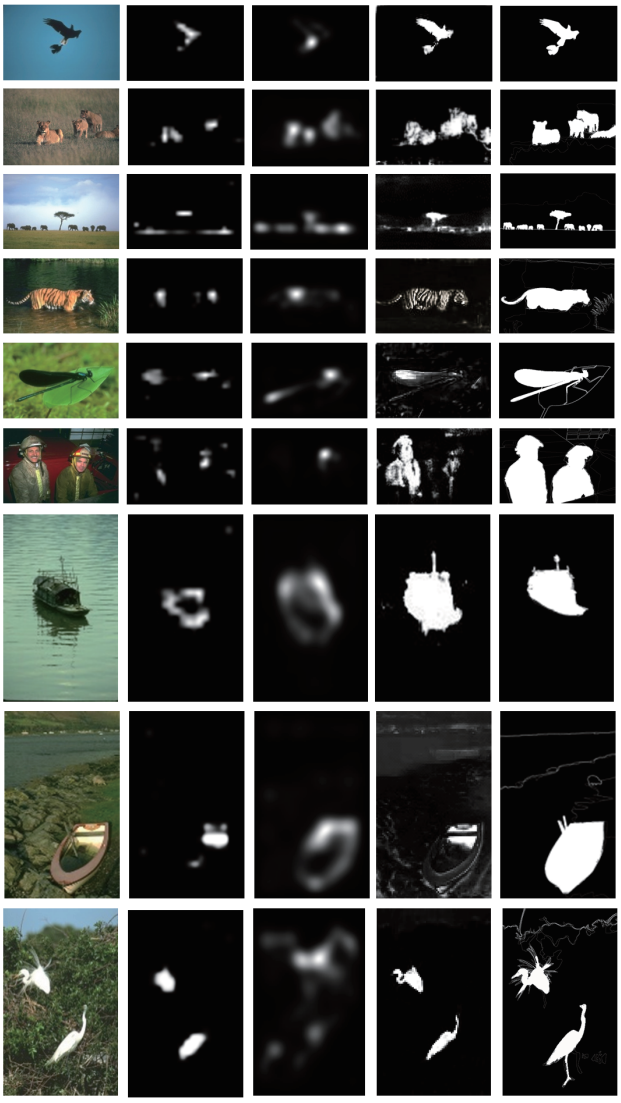
[7] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. In *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pages 2376–2383, 2010. 419, 422

[8] D. Grangier and S. Bengio. A discriminative kernel-based approach to rank images from text queries. *IEEE Trans. Pattern Anal. Machine Intell.*, 30(8):1371–1384, Aug. 2008. 420

[9] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Proc. Conf. Neural Inf. Proc. Systems*, pages 545–552, 2006. 418

[10] X. Hou and L. Zhang. Dynamic visual attention: searching for coding length increments. In *Proc. Conf. Neural Inf. Proc. Systems*, pages 681–688, 2008. 417, 422

[11] X. D. Hou and L. Q. Zhang. Saliency detection: A spectral residual approach. In *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pages 1–8, 2007. 422



(a) Input (b) Itti98 (c) Hou08 (d)Our extrinsic (e)Ground truth

Figure 9. Further comparison of saliency estimation results for various methods evaluated on BSDB data set against the extrinsic approach for $k = 20$.

[12] L. Itti and P. Baldi. Bayesian surprise attracts human attention. In *Proc. Conf. Neural Inf. Proc. Systems*, 2005. 418

[13] L. Itti and P. Baldi. A principled approach to detecting surprising events in video. In *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pages 631–637, 2005. 418

[14] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Machine Intell.*, 20(11):1254–1259, 1998. 417, 418, 422

[15] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurbiology*, 4:219–227, 1985. 418

[16] M. A. Kunar and J. M. W. S. J. Flusberg. Contextual cueing by global features. *Percept Psychophys*, 68(7):1204–1216, Oct. 2006. 418

[17] O. le Meur, P. le Callet, D. Barba, and D. Thoreau. A coherent computational approach to model bottom-up visual attention. *IEEE Trans. Pattern Anal. Machine Intell.*, 28(5):802–817, May 2006. 418

[18] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman. SIFT flow: Dense correspondence across different scenes. In *Proc. European Conf. Computer Vision*, pages III: 28–42, 2008. 419, 420, 421

[19] T. Liu, J. Sun, N. N. Zheng, X. Tang, and H. Y. Shum. Learning to detect a salient object. In *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pages 1–8, 2007. 418, 419, 421

[20] L. Marchesotti, C. Cifarelli, and G. Csurka. A framework for visual saliency detection with applications to image thumbnailing. *Proc. IEEE Int. Conf. Computer Vision*, 2009. 417, 418, 419

[21] D. R. Martin, C. C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. IEEE Int. Conf. Computer Vision*, pages II: 416–423, 2001. 418, 421

[22] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson. Top-down control of visual attention in object detection. In *Proc. IEEE Int. Conf. Image Processing*, pages 253–256, 2003. 418, 419

[23] B. H. Sotelo, A. Oliva, and A. B. Torralba. Human learning of contextual priors for object search: Where does the time go? In *Attention and Performance in Computational Vision*, pages III: 86–86, 2005. 418

[24] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Machine Intell.*, 30(11):1958–1970, Nov. 2008. 419

[25] A. Torralba, R. Fergus, and Y. Weiss. Small codes and large image databases for recognition. In *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pages 1–8, 2008. 419, 420

[26] A. Torralba, A. Oliva, M. Castelhano, and J. M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search. *Psychological Review*, 113(4):766–786, Oct. 2006. 419

[27] A. M. Triesman and G. Gelade. A feature-integration theory of attention. *ACM Computing Surveys Cognitive Psychology*, 12:97–136, 1980. 418

[28] J. K. Tsotsos, S. M. Culhane, W. Y. Wai, Y. H. Lai, N. Davis, and F. Nuflo. Modeling visual-attention via selective tuning. *Artif. Intell.*, 78(1-2):507–545, Oct. 1995. 418

[29] W. Wang, Y. Wang, Q. Huang, and W. Gao. Measuring visual saliency by site entropy rate. In *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pages 2368–2375, 2010. 418

[30] X.-J. Wang, L. Zhang, M. Liu, Y. Li, and W.-Y. Ma. ARISTA - image search to annotation on billions of web photos. In *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pages 2987–2994, 2010. 419

[31] J. Yuen and A. Torralba. A data-driven approach for event prediction. In *Proc. European Conf. Computer Vision*, volume 6312, pages 707–720, 2010. 419