A Proto-Object-Based Computational Model for Visual Saliency

Yanulevskaya, V.; Uijlings, J.; Geusebroek, J.-M.; Sebe, N.; Smeulders, A.

DOI
10.1167/13.13.27

Publication date
2013

Document Version
Final published version

Published in
Journal of Vision

License
CC BY-NC-ND

Link to publication

Citation for published version (APA):
Yanulevskaya, V., Uijlings, J., Geusebroek, J-M., Sebe, N., & Smeulders, A. (2013). A Proto-Object-Based Computational Model for Visual Saliency. *Journal of Vision*, *13*(13), [27]. https://doi.org/10.1167/13.13.27

A proto-object-based computational model for visual saliency

Victoria Yanulevskaya Department of Information Engineering and Computer Science, University of Trento, Italy

Jasper Uijlings Department of Information Engineering and Computer Science, University of Trento, Italy

Jan-Mark Geusebroek Intelligent Systems Lab Amsterdam, University of Amsterdam, the Netherlands

Nicu Sebe Department of Information Engineering and Computer Science, University of Trento, Italy

Arnold Smeulders Intelligent Systems Lab Amsterdam, University of Amsterdam, the Netherlands

State-of-the-art bottom-up saliency models often assign high saliency values at or near high-contrast edges, whereas people tend to look within the regions delineated by those edges, namely the objects. To resolve this inconsistency, in this work we estimate saliency at the level of coherent image regions. According to object-based attention theory, the human brain groups similar pixels into coherent regions, which are called *proto-objects*. The saliency of these proto-objects is estimated and incorporated together. As usual, attention is given to the most salient image regions. In this paper we employ state-of-the-art computer vision techniques to implement a proto-object-based model for visual attention. Particularly, a hierarchical image segmentation algorithm is used to extract proto-objects. The two most powerful ways to estimate saliency, rarity-based and contrast-based saliency, are generalized to assess the saliency at the proto-object level. The rarity-based saliency assesses if the proto-object contains rare or outstanding details. The contrast-based saliency estimates how much the proto-object differs from the surroundings. However, not all image regions with high contrast to the surroundings attract human attention. We take this into account by distinguishing between external and internal contrast-based saliency. Where the external contrast-based saliency estimates the difference between the proto-object and the rest of the image, the internal contrast-based saliency estimates the complexity of the proto-object itself. We evaluate the performance of the proposed method and its components on two challenging eye-fixation datasets (Judd, Ehinger, Durand, & Torralba, 2009; Subramanian, Katti, Sebe, Kankanhalli, & Chua, 2010). The results show the importance of rarity-based and both external and internal contrast-based saliency in fixation prediction. Moreover, the comparison with state-of-the-art computational models for visual saliency demonstrates the advantage of proto-objects as units of analysis.

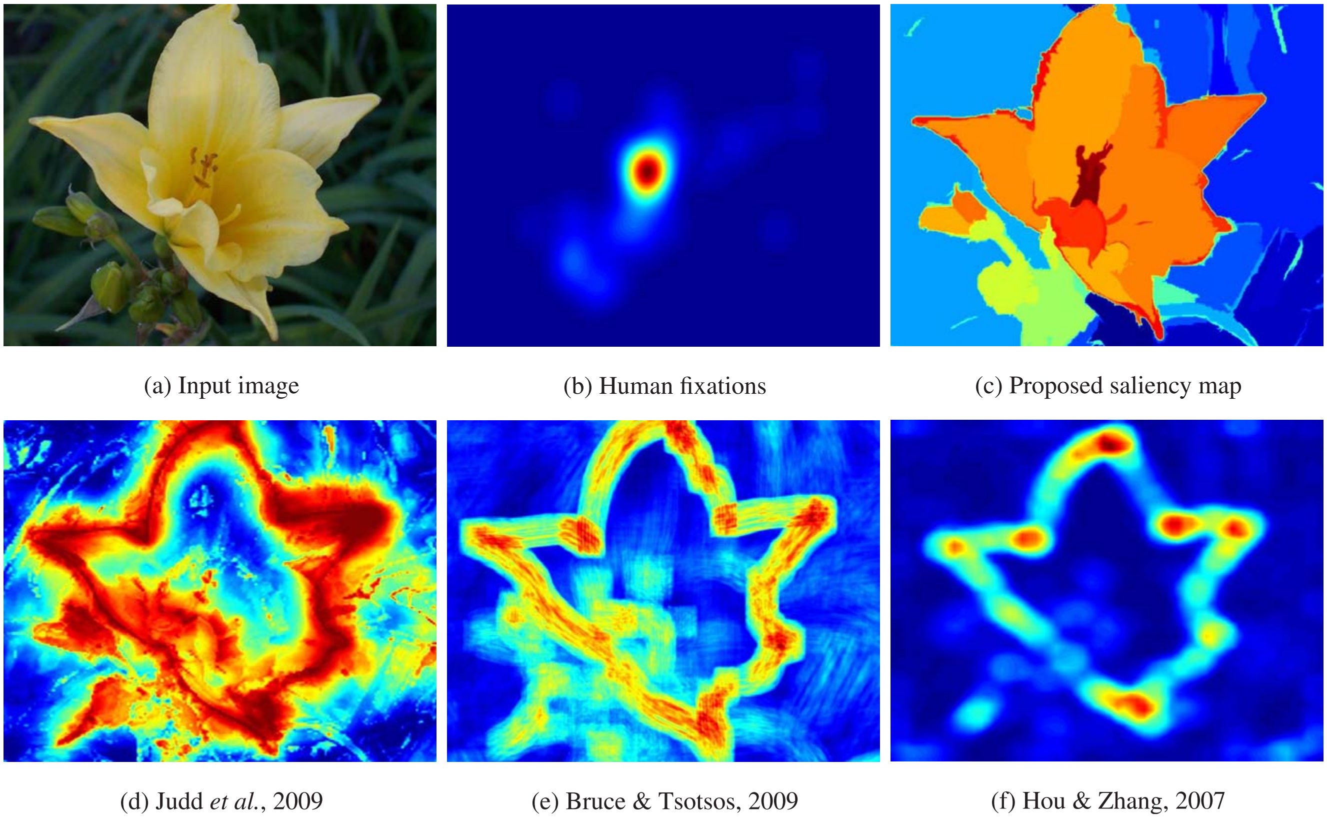
Introduction

To compensate for the difference in the acuity between the central and peripheral vision, people move their eyes three to four times a second, over 150,000 times each day. To sample the environment efficiently, the eye movements are influenced by the scene content and the goal of the observer. Thus, certain areas in the scene attract more fixations than others. Recently, a lot of progress has been made towards the full understanding of the mechanisms that underlie this visual sampling. It has resulted in many successful computational models for visual attention, such as Bruce and Tsotsos (2009), Gao, Mahadevan, and Vasconcelos (2008), Hou and Zhang (2007), Itti, Koch, and Niebur (1998), Judd et al. (2009), and Zhao and Koch (2011), which have achieved good performance in predicting human eye movements when viewing images.

The dominant computational models for visual attention are based on the estimation of scene saliency,

Citation: Yanulevskaya, V., Uijlings, J., Geusebroek, J.-M., Sebe, N., & Smeulders, A. (2013). A proto-object-based computational model for visual saliency. *Journal of Vision*, *13*(13):27, 1–19, http://www.journalofvision.org/content/13/13/27, doi:10.1167/13.13.27.

doi: 10.1167/13.13.27 Received February 8, 2012; published November 26, 2013 ISSN 1534-7362 © 2013 ARVO


(a) Input image (b) Human fixations (c) Proposed saliency map

(d) Judd *et al.*, 2009 (e) Bruce & Tsotsos, 2009 (f) Hou & Zhang, 2007

Figure 1. While people tend to look within an object (b), state-of-the-art computational models for visual attention often highlight the parts of an object with high contrast that mostly corresponds to object borders (d–f). We propose to resolve this by measuring the saliency at the proto-object level (c). Note that red values in saliency maps represent higher saliency, while blue values mean lower saliency.

where saliency is a visual uniqueness or rarity which makes some areas stand out from the surrounding and immediately grab attention (Treisman & Gelade, 1980). Usually, saliency is measured at the pixel level using image features like intensity, color, gradient, edges, and boundaries (Baddeley & Tatler, 2006; Itti et al., 1998; Zhang, Tong, Marks, Shan, & Cottrell, 2008). Therefore, as illustrated in Figure 1, the existing saliency maps tend to highlight scene areas of high contrast. However, such high-contrast edges are often located on the boundaries of objects, whereas people tend to look inside objects (Einhäuser, Spain, & Perona, 2008; Nuthmann & Henderson, 2010).

In this work, we move towards attributing saliency to objects in an image. Particularly, we follow an object-based attention theory (Duncan, 1984; Egly, Driver, & Rafal, 1994; Vecera & Farah, 1994). According to this theory, at the early pre-attentive stage, the visual system pre-segments a complex scene into proto-objects, where a proto-object is a coherent region that approximates an object, a part of object, or a group of objects (Beck, 1966; Julesz, 1981; Julesz & Bergen, 1983; Rensink, 2000). Then saliency is estimated at the level of proto-object, and more salient proto-objects attract more attention.

The difficult part of building an object-based computational model is to separate objects from the background. The methods that require precise object locations (Einhäuser et al., 2008; Nuthmann & Henderson, 2010; Vincent, Baddeley, Correani, Troscianko, & Leonards, 2009) are constrained by manual object outlining because accurate automatic segmentation for generic objects requires a priori knowledge and is still beyond the current automatic techniques. Moreover, accurate automatic segmentation implies that all objects are already recognized, whereas attention is believed to start acting before object recognition (Walther & Koch, 2006). To relax this requirement for precise object delineation, we consider proto-objects as units of analysis. Recent research in object detection (van de Sande, Uijlings, Gevers, & Smeulders, 2011) and salient object detection (Yanulevskaya, Uijlings, & Geusebroek, 2013) has shown that approximations of object locations can be successfully used in practical applications. Therefore, we adapt the hierarchical image segmentation used in van de Sande et al. (2011) to extract proto-objects of an image.

Based on the intuition that visually salient objects attract more attention than nonsalient objects, the importance of a proto-object can be estimated by assessing how much the proto-object "pops out" from the scene. From the literature, it is known that an image region pops out in two cases: (a) when it contains rare or outstanding details (Alexe, Deselaers, & Ferrari, 2010; Bruce & Tsotsos, 2009; Yanulevskaya et al., 2013; Zhang et al., 2008), and (b) when it differs from the surroundings (Cheng, Zhang, Mitra, Huang, & Hu, 2011; Duncan & Humphreys, 1989; Itti et al., 1998; Kienzle, Franz, Scholkopf, & Wichmann, 2009; Krieger, Rentschler, Hauske, Schill, & Zetzsche, 2000; Liu et al., 2010; Reinagel & Zador, 1999; Rosenholtz, 1999). However, not all image regions that differ from the surroundings attract human attention. For example, a piece of sky may differ considerably from the rest of an image and still people do not look at it. This happens because, in general, the sky has a uniform structure without salient details. We take this into account by distinguishing external and internal contrast-based saliency, where the external contrast-based saliency estimates the difference between the image region and the rest of image, and the internal contrast-based saliency estimates the complexity of the proto-object itself. We make all measurements at the level of proto-objects, which allows for incorporating the notion of an object directly into the saliency estimation process. Thereby, the important image regions are highlighted in their entirety as it is illustrated in Figure 1.

In this paper we are building on the approach for salient object detection proposed in Yanulevskaya et al. (2013). In Yanulevskaya et al. (2013), the method selects the most salient proto-object that captures a complete object. In the current work, we investigate the link between the saliency of proto-objects and the way people look at images. Particularly, the hypothesis is that the more salient a proto-object is, the more fixations it will attract. Thus in this work, the saliency of all proto-objects is incorporated into a single saliency map, which predicts where people look while observing an image. Moreover, in comparison with Yanulevskaya et al. (2013), we extend the measurement of the contrast-based saliency by introducing the external and internal contrast of a proto-object.

Related work

Many successful computational models for visual saliency have been proposed recently (such as Bruce & Tsotsos, 2009; Cerf, Frady, & Koch, 2009; Gao et al., 2008; Harel, Koch, & Perona, 2007; Itti et al., 1998; Judd et al., 2009; Kadir & Brady, 2001; Kienzle et al., 2009; Renninger, Coughlan, Verghese, & Malik, 2005; Rosenholtz, 1999; Tatler, Baddeley, & Gilchrist, 2005; Zhao & Koch, 2011). These models range from biologically plausible to pure computational and the combination of the two. Itti et al. (1998) proposed a model inspired by the primate visual system. From what is known to be extracted in early cortical areas, they constructed a saliency map by combining color, contrast, and orientation features at various scales. They implemented a center-surround operation by taking the difference of feature-specific maps at two consecutive scales. The result for each feature is normalized, yielding three conspicuity maps. The overall saliency map is a linear combination of these conspicuity maps. Their influential approach has set a standard in saliency prediction. Rosenholtz (1999) suggested that in visual search the saliency of a target depends on deviation of its feature values from the average statistics of the image. In other words, statistical outliers are salient. Bruce and Tsotsos (2009) formulated this principle in terms of the information theory. They proposed a computational model where saliency is calculated as Shannon's self-information. Intuitively, image locations with unexpected content in comparison with their surrounding are more informative, and thus salient. Hou and Zhang (2007) examined images in the spectral domain. They demonstrated that the statistical singularities in the spectrum domain correspond to regions in the image that differ from the surrounding. Thus, the spectral residual is indicative to the saliency. Judd et al. (2009) combined machine learning techniques with successful saliency models and high-level image information. They jointly considered features from Itti et al. (1998), Oliva and Torralba (2001), and Rosenholtz (1999), the steerable pyramid filters (Simoncelli & Freeman, 1995), location of the horizon, locations of some objects like people and cars, and position in relation to the center of the image. Then a classifier is trained on recorded eye movements to combine all features in an optimal way. We will demonstrate that the proposed method consistently outperforms the approaches of Bruce and Tsotsos (2009) and Hou and Zhang (2007) and reaches the level of Judd et al. (2009), whereas our method does not require any learning.

Einhäuser et al. (2008) investigated the role of objects in visual attention. In their experiments they manually segmented images to localize objects. Then, in addition to eye-movement recording, they asked subjects to name the most interesting objects within an image. The authors demonstrated that, weighted by the recall frequency, locations of the objects predict eye-fixations better than the standard method by Itti et al. (1998). Our method does not require the manual annotation, instead we propose a way to automatically extract and estimate saliency of proto-objects.

The correlation between the significance of an object and its saliency has been demonstrated by Elazary and Itti (2008). In their experiments, Elazary and Itti (2008) considered the LabelMe database (Russell, Torralba, Murphy, & Freeman, 2008), which contains images with some objects manually segmented by a large population of users. Importantly, users themselves decided which objects they would like to annotate. Therefore, it is hypothesized that people segmented objects that attracted their attention. The authors demonstrated that the high peaks of the saliency map (Itti et al., 1998) coincide with the segmented objects. This implies that objects that attract attention tend to contain visually outstanding details. Therefore, in our work, we use saliency of proto-objects to predict where people focus their attention.

Walther and Koch (2006) proposed a biologically plausible model of automatically forming and attending to proto-objects in natural scenes. They considered the pixel level saliency map of Itti et al. (1998) and determined the spatial extent of its peaks as proto-objects. Particularly, an extracted proto-object is a set of neighboring pixels with saliency above a certain threshold. In such an approach, proto-objects are determined mostly by the structure of the pixel-based saliency map whereas the information about image structures is not taken into account explicitly. In this paper, the proto-objects are extracted directly from the image by hierarchical segmentation. Therefore, the arrangement of objects of an image determines the shape of proto-objects.

Recently several computational proto-object based models for visual attention (Orabona, Metta, & Giulio, 2007; Wischnewski, Belardinelli, Schneider, & Steil, 2010) have addressed a question "Where to look next?" These models are designed to predict the next saccade target. Therefore, they are looking for the most salient proto-object given the current fixation location. In our work, we estimate the overall importance of each proto-object to predict the distribution of fixation locations. In Wischnewski et al. (2010), proto-objects are extracted at multiple scales but are approximated by ellipses. In Orabona et al. (2007), proto-object are represented as segments with a precise boundary instead of ellipses, yet only a single scale is considered. Our proto-objects capture favorable aspects of both Wischnewski et al. (2010) and Orabona et al. (2007): We hierarchically segment an image to obtain proto-objects with a precise boundary at multiscales. Finally, we employ more advanced visual features for both the proto-object generation and the saliency estimation.

Tatler, Hayhoe, Land, and Ballard (2011) discussed some limitations of saliency-based models to explain eye guidance in natural vision. First of all, as have been shown by many (Einhäuser, Rutishauser, & Koch, 2008; Foulsham & Underwood, 2008; Henderson, Brockmole, Castelhano, & Mack, 2007; Underwood, Foulsham, Van Loon, Humphreys, & Bloyce, 2006), image structures that pop out are not always looked at, especially when the observer has a specific task in mind. Moreover, Tatler et al. (2011) pointed out that recording of the eye-movements in laboratory settings in front of a static screen does not fully reflect the complexity of visual behavior in a more natural dynamic environment. Nevertheless, Tatler et al. (2011) agreed that saliency does play a role in allocation of eye movements and that saliency-based models for visual attention can provide explanations of the way people look around. In this paper we aim to improve the current saliency models.

In this paper we make the following contributions: (a) We incorporate the notion of object into saliency measurements by considering a proto-object as a unit of analysis, where proto-objects are extracted automatically. (b) To extract proto-objects we segment the image, rather than its derivative pixel-level saliency map. (c) Three types of saliency of proto-objects, i.e., rarity-based, external, and internal contrast-based saliency, are considered to predict where people will look in an image. We demonstrate that the proposed method compares favorably with the state-of-the-art models in saliency prediction on two challenging eye-fixation datasets.

Saliency map based on proto-objects

A proto-object is a coherent image region that, by the visual coherence in most objects in the world, roughly corresponds to part of an object, a complete object, or a group of objects. Hence, an object may consist of several small proto-objects approximating its parts and, at the same time, be part of a larger proto-object that contains a group of objects. Therefore, proto-objects are organized in a hierarchical way, which suggests that they can be extracted from an image using a hierarchical segmentation. Following Yanulevskaya et al. (2013), we adapted the hierarchical image segmentation used in van de Sande et al. (2011) to extract proto-objects. Afterwards the saliency of all segments is estimated and combined into the final saliency map.

Proto-objects extraction

We use the graph-based hierarchical image segmentation (van de Sande et al., 2011) to obtain a set of proto-objects. It starts with an over-segmentation of an image using the popular algorithm by Felzenszwalb and Huttenlocher (2004). Then, the image is represented as a graph, where nodes are segments and edges represent similarities between neighboring segments.

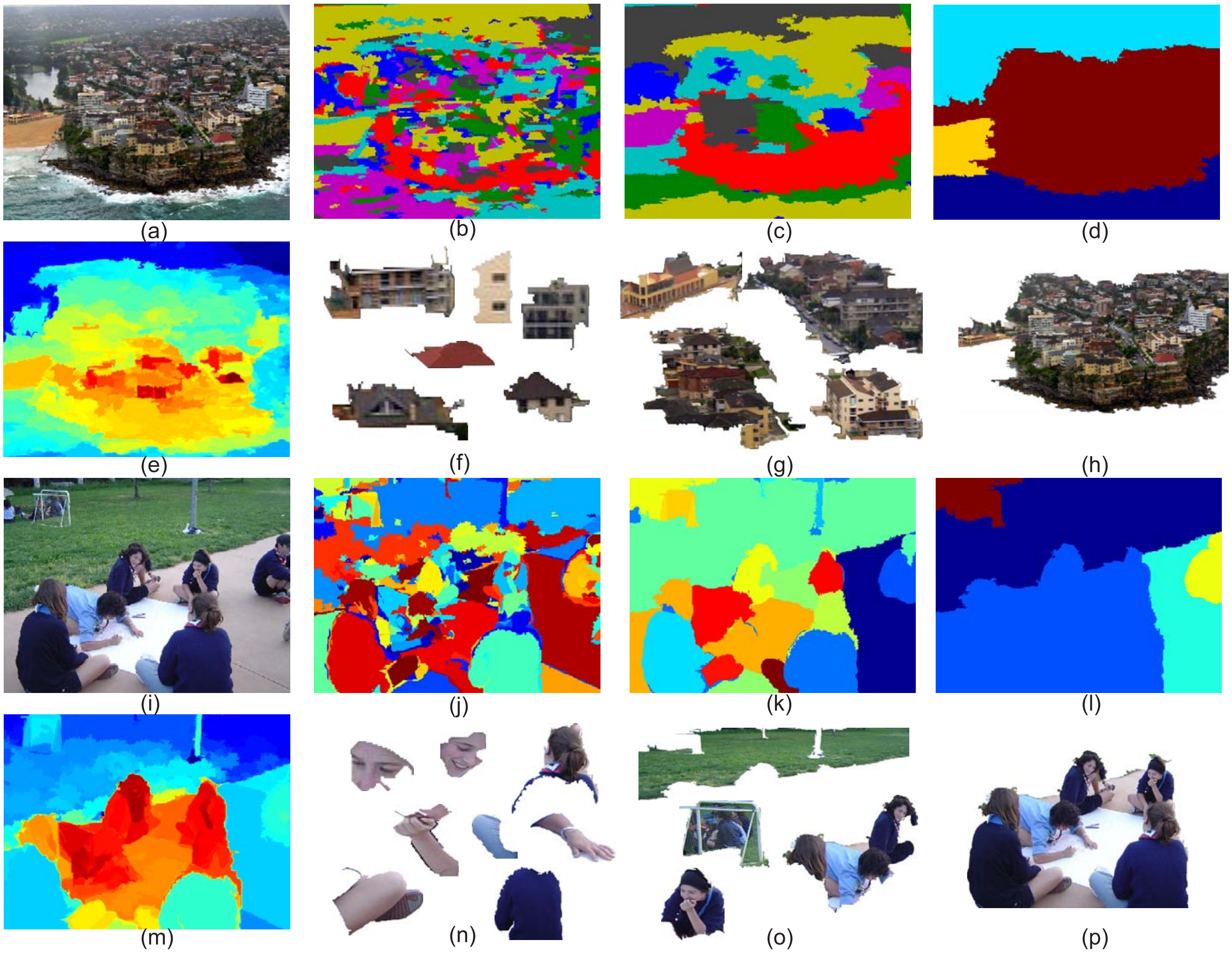


Figure 2. Two examples of hierarchy of proto-objects: For two images (a) and (i) we start with the initial over-segmentation (b) and (j). The proto-objects at this scale correspond to distinct parts of objects, for example to roofs and walls of a building in (f), or to different body parts in (n). As neighboring segments are merged according to their similarity, see (c) and (k), we obtain proto-objects that correspond to the objects and small groups of similar objects: for example buildings and districts in (g), or individual persons and groups of people who sit next to each other in (o). As we proceed to merge segments in (d) and (l), the proto-objects extend to the main elements of the images, for example to the whole city in (h), or to the largest group of people in (p). Therefore, such hierarchy of proto-objects allows us to incorporate into the saliency maps (e) and (m) characteristics of all image structures from small object details to large groups of objects.

To create a hierarchy, the edge with the highest similarity is iteratively selected and the corresponding segments are merged. The similarities between this newly merged segment and its neighbors are updated. This process is repeated until the whole image becomes a single segment. Figure 2 gives an example of the hierarchy of proto-objects.

In this work we enrich the set of similarity measurements used in hierarchical segmentation in comparison with the original implementation of van de Sande et al. (2011). Van de Sande et al. (2011) used texture-based and size-based similarities. We also include color-based similarity together with spatial relationship between segments. To capture the texture of a segment, gradient responses in four directions (0°, 45°, 90°, 135°) are calculated. Then, the texture-based similarity $S_{texture}(a,b)$ is estimated as the histogram intersection of these gradient responses. The histogram intersection is defined as follows:

$$D(h,h') = 1 - \sum_{k=1}^{K} \min(h_k, h'_k), \qquad (1)$$

where h and $h^{'}$ are two histograms of length K.

The size similarity $S_{size}(a,b)$ forces smaller regions to be merged first. As a consequence, segments of similar scales appear at the same level of hierarchy. It is defined as

$$S_{size}(a,b) = \frac{|I| - |a| - |b|}{|I|}, \tag{2}$$

where I stands for the whole image and $|x|$ is the number of pixels in segment x.

Furthermore, to estimate the color-based similarity between two segments $S_{colour}(a,b)$, we calculate a color histogram of each segment and the histogram intersection of them. With spatial relationship between two segments $S_{enclosed}(a,b)$ we capture to which extent one segment is enclosed in another. This closes holes inside segments and encourages convexness. If $Bn(a)$ is the number of boundary pixels of a, and $Bn(a) < Bn(b)$,

$$S_{enclosed}(a,b) = \frac{Br(a,b)}{Bn(a)}, \tag{3}$$

where $Br(a, b)$ counts the number of pixels of segment a that touch segment b. Therefore, if a is completely enclosed by b, then $S_{enclosed}(a,b)$ is one.

Thus, the final similarity between two segments is calculated as a linear combination of four measurements:

$$S(a,b) = S_{texture}(a,b) + S_{size}(a,b) + S_{colour}(a,b) + S_{enclosed}(a,b). \tag{4}$$

Proto-object saliency estimation

To assess the saliency of a proto-object we estimate how much it pops out. An image region pops out when it differs from the surroundings (Cheng et al., 2011; Itti et al., 1998; Kienzle et al., 2009; Krieger et al., 2000; Liu et al., 2010; Reinagel & Zador, 1999; Rosenholtz, 1999) and when it contains rare or outstanding details (Alexe et al., 2010; Bruce & Tsotsos, 2009; Yanulevskaya et al., 2013; Zhang et al., 2008). In our method contrast-based and rarity-based saliency of proto-objects are combined.

Contrast-based saliency

A difference in appearance from the surrounding is both an object characteristic and an indication of saliency. Therefore, by measuring the contrast of proto-objects to their surroundings we estimate the *saliency* and at the same time we encourage proto-objects that approximate image objects more accurately. Recently, Cheng et al. (2011) proposed a global contrast-based method to estimate saliency of a region. They segmented an image into the set of nonintersecting regions. Then they assessed saliency of a region by estimating the contrast between this region and all other regions in the image. In this paper, we have adapted their approach for hierarchically overlapping regions in the following way. Let P be an initial image segmentation. We calculate the saliency of a proto-object a as the average of color contrasts between this proto-object a and all surrounding segments from P. We call this *external contrast*. Note that P corresponds only to the first finest level of the proto-object hierarchy. We cannot compare a to all surrounding proto-objects throughout the hierarchy as some image parts are covered by a larger number of proto-objects than others do. This would distort the measurements.

Intuitively, the difference between neighboring proto-objects is more important than between remote ones. Therefore, the contrast should be weighted by the spatial distance between proto-objects. Thus, the weights are calculated as a Gaussian function $\exp(-\frac{D(a,b)}{\sigma^2})$ of the Euclidean distance between the centroids of corresponding proto-objects $D(a, b)$. The parameter σ allows the algorithm to control the contribution of remote segments to saliency estimation. Furthermore, the contrast to larger segments influences proto-object saliency more than the contrast to smaller segments. Therefore, the contrast from proto-object a to proto-object b is weighted by the number of pixels $|b|$ within b. To summarize, the final equation for the proto-object saliency based on the external contrast looks as follows:

$$Sal_{external}(a) = \frac{\sum p_i \in P \setminus aC(a,p_i)\exp(-\frac{D(a,p_i)}{\sigma^2})|p_i|}{|P| - |a|}, \tag{5}$$

where p_i are segments from the initial segmentation P and $C(a,p_i)$ is the contrast between proto-objects a and p_i. The number of pixels outside a $|P| - |a|$ is a normalization factor.

The saliency of a proto-object also depends on the complexity of the proto-object itself (Kadir & Brady, 2001; Renninger, Verghese, & Coughlan, 2007). A bright sky may differ a lot from the rest of an image, and yet people usually do not look at it. This happens because the sky mostly consists of a uniform smooth area without distinctive details. Therefore, to estimate the complexity of a proto-object we compare its parts with each other. Particularly, we estimate the average difference between all segments within a proto-object, which we call *internal contrast*. Here, we again consider only the initial segmentation P. As in a case of external contrast, some image parts are covered by a larger number of proto-objects than others do, which would distort the measurements. The internal contrast is weighted in the same way as the external one. Thus, the saliency of a proto-object based on the internal contrast is defined as follows:

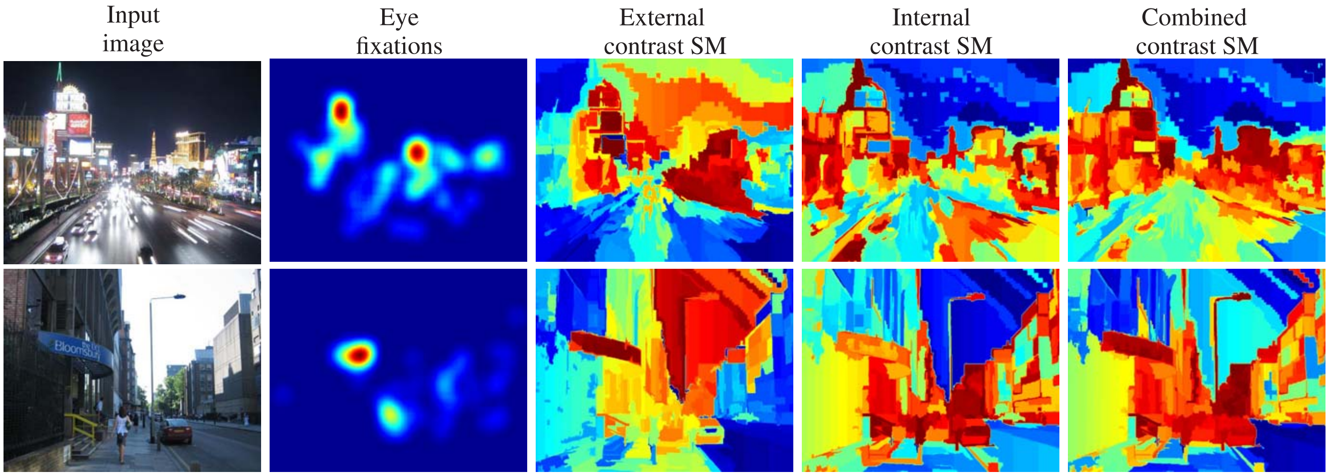


Figure 3. An example of contrast-based saliency maps. People generally do not attend homogeneous areas like sky and roads even when such areas have a very contrasting color, which is why they are highlighted by the external contrast-based saliency map. We account for this phenomenon by combining external and internal contrast-based saliency.

$$Sal_{internal}(a) = \frac{1}{n}\sum_{p_i \in P\cap a} \frac{\sum p_j \in P\cap a C(p_i,p_j)\exp(-\frac{D(p_i,p_j)}{\sigma^2})|p_j|}{|a| - |p_i|}, \qquad (6)$$

where n is a number of segments in $P\cap a$. Note that for $i = j$, $C(p_i, p_i) = 0$.

We linearly combine external and internal contrasts into the contrast-based saliency:

$$Sal_{contrast}(a) = Sal_{external}(a) + Sal_{internal}(a). \qquad (7)$$

Note that for now we consider that the internal and the external contrast-based saliency are equally important. In the Evaluation section we will investigate different weighting schemes.

To create a contrast-based saliency map we average the saliency of proto-objects calculated according to Equation 7 over the pixels they cover. Figure 3 illustrates several examples of external, internal, and combined contrast-based saliency maps. The saliency based on external contrast tends to highlight uniform regions like sky and water; however, the combination with saliency based on internal contrast effectively resolves this problem.

Rarity-based saliency

People tend to look at rarely occurring image structures (Bruce & Tsotsos, 2009; Rosenholtz, 1999; Zhang et al., 2008). Therefore, these structures should be highlighted. In contrast, frequently occurring patterns are typically part of the image background and are not fixated. They should be suppressed. To capture image structures we represent an image as a bag-of-visual words (Csurka, Dance, Fan, Willamowski, & Bray, 2004; Sivic & Zisserman, 2003), which is the state-of-the-art technique in object detection. In this representation every pixel of an image is associated with a rectangular patch around it. Then for each patch the Scale-Invariant Feature Transform (SIFT) descriptor (Lowe, 2004) is computed. Particularly, a patch is divided into 4 × 4 cells, where within each cell gradients in eight orientations are summed together. This representation efficiently captures both contours and texture of an image patch. Furthermore, the set of SIFT descriptors is quantized where the clusters are called *visual words*. Thereby, each pixel in an image is associated with a patch, where each patch of an image is mapped into a visual word, so that the whole image may be represented as a *bag-of-visual words* (Csurka et al., 2004; Sivic & Zisserman, 2003). As an image representation in terms of SIFT visual words is currently the most effective one in the object recognition task, this representation is believed to capture important object structures. Thus, the distribution of visual words within an image may be indicative for saliency. In our method, rare visual words are considered salient. Particularly, inspired by the information maximization approach (Bruce & Tsotsos, 2009), we calculate the saliency of a pixel p_i as the self-information of the corresponding visual word w_{p_i}:

$$Sal_{pixel}(p_i) = -log(P(w_{p_i})), \qquad (8)$$

where $P(w_{p_i})$ is the probability of a visual word w_{p_i} defined relative to each image.

Equation 8 defines saliency at the pixel level. To extend it to the level of proto-objects, we average the saliency of all pixels within a proto-object:

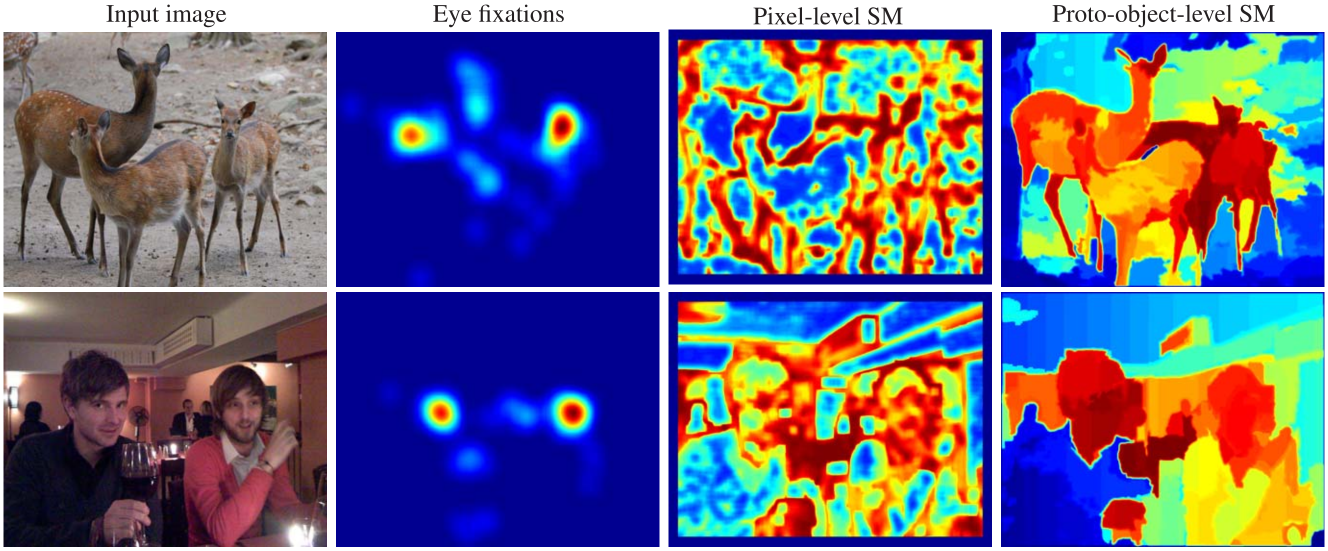


Figure 4. Examples of rarity-based saliency map at pixel and proto-object levels. While people tend to fixate inside image regions such as animal snouts and human faces, pixel level saliency maps highlight mostly objects edges while ignoring inner parts. By spreading saliency over proto-objects, we effectively redistribute it. As a result, on the proto-object-based saliency maps (most right column) objects are highlighted more uniformly.

$$Sal_{rarity}(a) = \frac{1}{|a|}\sum_{p_i \in a} Sal_{pixel}(p_i), \qquad (9)$$

where $|a|$ is the number of pixels within a proto-object a. As in the previous section, to create a rarity based proto-object saliency map, we average the saliency of proto-objects over the pixels they cover. Figure 4 illustrates the difference between the rarity-based saliency at pixel and proto-object levels. While pixel-based measurements highlight mostly object edges, Equation 9 smoothes the saliency over proto-objects, thereby effectively highlighting entire objects.

Combined proto-object-based saliency map

In this section we describe the final saliency map. As contrast-based saliency employs color information, and rarity-based saliency is based on texture information, these two measurements are complementary to one another. Therefore, we combine the contrast-based and rarity-based saliency into the final saliency measurement:

$$Sal_{proto-object}(a) = Sal_{contrast}(a) + Sal_{rarity}(a). \qquad (10)$$

Based on this equation, we estimate the saliency for all proto-objects. Then, because proto-objects have a hierarchical structure and may be overlapping, we average proto-object saliency over the pixels they cover. The resulting final saliency map predicts where people look while investigating an image. In the Evaluation section we will investigate different weights for each saliency term.

Implementational details

Proto-object extraction

In the same way as in Yanulevskaya et al. (2013), we run Felzenszwalb and Huttenlocher's algorithm (2004) to over-segment an image. However, in this paper we segment the image only once with the following settings: RGB color space, the smoothing parameter is 0.8, and the scale parameter is 100. Then, from this initial over-segmentation, a hierarchical segmentation is generated as described in Section "Proto-object extraction." To estimate texture similarity, we calculate gradient responses with $\sigma = 0.8$. We filter out segments that are smaller than 30 × 30 pixels (the average image size is 1024 × 768), and we remove segments that overlap more than 70% with other segments. This process results in about 1,000 proto-objects per image.

Contrast-based saliency

We estimate the contrast between two proto-objects $C(a, p_i)$ as chi-square distance between their color histograms. Specifically, we use three-dimensional histograms in the $L{*}a{*}b{*}$ color space with 6 × 6 × 6 bins. Furthermore, as recommended in Cheng et al. (2011), we set the parameter σ^2 from Equations 5 and 6 to 0.4, where pixel coordinates are normalized to [0, 1].

Rarity-based saliency

We calculate visual words using the framework of Uijlings, Smeulders, and Scha (2009). To be precise, the standard intensity-based SIFT descriptor is used, which covers an image patch of 24 × 24 pixels. The histogram of local gradient directions is computed at a single scale ($\sigma = 0.8$) on a regular image grid. This SIFT implementation is not rotation invariant. In order to retain contrast information, SIFT is not normalized to a unit vector. To create a visual vocabulary, 250,000 randomly selected SIFT descriptors are quantized into 4,096 clusters using K-means. As SIFT calculation involves image gradients, it is impossible to reliably extract visual words at the image borders. In order to avoid artifacts, as suggested in Bruce and Tsotsos (2009), we ignore the outer 24 pixels of the image borders. This is equal to the width of the region from which a visual word is extracted. As a side effect, saliency of peripheral proto-objects touching the border is slightly reduced. However, this effect matches the tendency of observers to attend the central part of an image, which has been frequently reported in the literature (Bruce & Tsotsos, 2009; Judd et al., 2009; Tatler, 2007; Zhang et al., 2008).

Evaluation

We test the proposed proto-object-based method on two recent eye-movement datasets: MIT (Judd et al., 2009) and NUSEF (Subramanian et al., 2010), where the task is to predict where people fixate while observing images. Strictly speaking, attentional and gaze shifts do not always coincide: In some specific cases the attentional focus can be directed to the new target without accompanied eye-movements (Horowitz, Fine, Fencsik, Yurgenson, & Wolfe, 2007; Kelley, Serences, Giesbrecht, & Yantis, 2008). However in everyday viewing conditions, they are tightly linked (Posner, 1980).

Evaluation method

The standard evaluation in eye-fixation prediction task (Bruce & Tsotsos, 2009; Judd et al., 2009; Tatler et al., 2005) is to calculate the area under the receiver operating characteristic (ROC) curve. In this case, a set of binary maps is generated by thresholding the evaluated saliency map. Then, the true positive and false positive rates are calculated for each binary map. The true positive rate is the fraction of fixated pixels above the threshold, while the false positive rate is the fraction of nonfixated pixels above the threshold. The ROC curve depicts the tradeoff between the true positive and false positive rates over various thresholds, where the area under the ROC curve (AUC) is regarded as an indication of an over all accuracy. For the perfect saliency map the AUC is 1, and for a random saliency map the AUC is 0.5.

To calculate the performance of a computational model we estimate how well the fixation locations of one subject are predicted by the generated saliency map. To compare the quality of computational models with respect to a human performance, we also compute the intersubject predictive power. Similarly to Judd et al. (2009), we estimate how well fixation locations of one subject are predicted by the fixation map generated based on eye-movements of the rest of the subjects. Where the fixation map is a convolution of fixation locations with a fovea-sized two-dimensional Gaussian kernel ($\sigma = 1°$, i.e., 30 pixels). We calculate AUCs for all subjects and images in the dataset. The average results are reported.

Eye-fixation datasets

We consider two eye-movement datasets: MIT (Judd et al., 2009) and NUSEF (Subramanian et al., 2010), both collected under the free-viewing task, i.e., subjects were asked to explore images without any specific task in mind. The MIT dataset by Judd et al. (2009) contains 1,003 images and eye-fixations of 15 subjects acquired over a period of 3 s. All images are randomly selected from Flickr and LabelMe (Russell et al., 2008). They have diverse appearances of everyday scenes ranging from landscapes and portraits to close-ups and graffiti. For this reason the dataset is representative and challenging. In our experiments we consider all recorded fixations. Figure 5 illustrates a number of images from the MIT dataset.

The NUSEF dataset by Subramanian et al. (Subramanian et al., 2010) contains 751 images and fixations of on average 24 subjects acquired over a period of 5 s. Apart from everyday live scenes, this dataset contains emotion-evoking images, nude depictions, and action scenes. Images are collected from various sources: Flickr, Photo.net, Google.com, and IAPS dataset (Lang, Bradley, & Cuthbert, 1999). In our experiments we consider all recorded fixations. Figure 6 shows representative images from the NUSEF dataset.

Results

We evaluate separately contrast-based saliency, rarity-based saliency, and their combination to analyze the contribution of each component of the proposed method. Furthermore, the proposed proto-object-based method is compared with the state-of-the-art approaches (Bruce & Tsotsos, 2009; Hou & Zhang, 2007; Itti et al., 1998; Judd et al., 2009) and with the intersubject variability. The original code provided by

Figure 5. Example of images from MIT dataset (Judd et al., 2009).

the authors was used to compute saliency maps of Bruce and Tsotsos (2009), Hou and Zhang (2007), Itti et al. (1998), and Judd et al. (2009). In the case of Hou and Zhang's method (2007), as suggested by the authors, images were rescaled to 64 × 64 pixels to calculate saliency map. The saliency maps were upscaled to the original size. In the case of the Judd et al. method (2009), for the MIT dataset we used the saliency maps provided by the authors. To generate saliency maps for NUSEF dataset, we used a trained model provided by the authors. The results are shown in Tables 1 and 2.

The internal contrast-based saliency alone achieves a moderate AUC of 0.689 on MIT dataset and 0.656 on NUSEF dataset, see Table 1. This type of saliency measures the difference between the proto-object itself and its parts. It allows the algorithm to filter out large nearly uniformly colored proto-objects, which usually belong to the background: grass, sky, or water, for example. Thus, the internal contrast-based saliency should be accompanied by the other types of saliency. The external contrast-based saliency estimates how much the proto-object varies from its surrounding. As it is shown in Table 1, the external contrast-based saliency alone has an AUC of 0.736 and 0.734 on MIT and NUSEF, respectively. Whereas the combined contrast-based saliency reaches an AUC of 0.748 and 0.746 on MIT and NUSEF, respectively.

The comparison of the rarity-based saliency at the pixel and proto-object levels illustrates the power of proto-objects as units of analysis. When the saliency is effectively spread over proto-objects, the performance rises from 0.744 to 0.778 on MIT and from 0.713 to 0.759 on NUSEF (see Table 1). The performance is boosted further when contrast- and rarity-based saliency are combined together: AUC of 0.785 and 0.770 on MIT and NUSEF, respectively. This indicates that contrast-based and rarity-based measurements are

Figure 6. Example of images from NUSEF dataset (Subramanian et al., 2010).

Type of saliency map	MIT	NUSEF
Internal contrast-based	0.689	0.656
External contrast-based	0.736	0.734
Combined contrast-based	0.748	0.746
Rarity-based at pixel level	0.744	0.713
Rarity-based at proto-object level	0.778	0.759
Combined proto-object-based	0.785	0.770

Table 1. Evaluation of each component of the proposed method. The internal contrast-based saliency does not show good performance, but combined contrast-based saliency and especially the rarity-based saliency at the proto-object level achieve good results. The combined proto-object-based saliency has the best performance.

Type of saliency map	MIT	NUSEF
(Bruce & Tsotsos, 2009)	0.735	0.706
(Hou & Zhang, 2007)	0.724	0.716
(Judd et al., 2009) (with CB)	0.815	0.817
(Judd et al., 2009) (without CB)	0.760	0.749
Proto-object-based (with CB)	0.823	0.839
Proto-object-based (without CB)	0.785	0.770
Intersubject variability	0.894	0.883

Table 2. Comparison of the proposed approach with the state-of-the-art computational models for visual saliency.

complimentary and both are important for saliency prediction.

Table 2 shows the performance of the state-of-the-art saliency models together with the intersubject variability. The proposed method outperforms saliency maps of Bruce and Tsotsos (2009) and Hou and Zhang (2007), whereas the method of Judd et al. (2009) demonstrates the best results on both MIT and NUSEF datasets. However, the method of Judd et al. (2009) has built in central bias, whereas other considered methods do not. It has been shown that the central part of an image, in general, attracts more eye fixations than the peripheral part (Buswell, 1935; Mannan, Ruddock, & Wooding, 1996; Parkhurst, Law, & Niebur, 2002; Tatler, 2007). Moreover, the Gaussian blob centered in the middle of the image usually shows excellent results that outperform automatic models of attention (Judd et al., 2009; Zhao & Koch, 2011). To compensate for the power of the central bias, we perform two additional experiments. First, we exclude the central bias from the method of Judd et al. (2009). Second, we include the central bias into our method. As can been seen in Table 2, in these more balanced settings, the proposed method has the best results: AUC of 0.785 versus 0.760 on MIT dataset and AUC of 0.770 versus 0.749 on NUSEF dataset when central bias is not taken into account by both methods and AUC of 0.823 versus 0.815 on MIT dadaset and AUC of 0.839 versus 0.817 on NUSEF dataset when central bias is integrated inside both methods.

Figures 7–9 provide visual comparisons of the proto-object-based saliency map with the saliency maps by Judd et al. (2009), Hou and Zhang (2007), and Bruce and Tsotsos (2009), and with human eye-fixations. To make a fair comparison, all saliency maps are considered without central bias. Not surprisingly, the advantage of the proposed method is most pronounced for images containing interesting objects. Figure 7 demonstrates that although some outstanding features might make an object interesting, people do not only fixate on the most salient details of the object. Instead, they tend to inspect the object more thoroughly. For example, in Figure 7, the most distinguishing detail of a dish, shown in the first row, is a flower on top of it, as is correctly captured by all saliency detectors. Nevertheless people, possibly attracted by the flower, examine the less outstanding parts of the dish as well, expanding the saliency of object details to the whole object. The proposed method succeeds in mimicking this behavior.

In some cases when an image contains a wireframe or a textured object on uniform background, all of the considered methods manage to highlight the whole object, see Figure 8. Therefore, for such images the performance of all methods is comparable.

However, we do not claim that the proposed approach explains eye-movements in all possible situations. The way people observe the scene is strongly affected by cognitive factors. Therefore modeling of top-down factors is necessary for a complete understanding of the gaze patterns. In this work, we concentrate on the bottom-up saliency. Thus our method is not designed for images with objects that are semantically important rather than visually salient. Some examples of such images are given in Figure 9. Another difficult case, which is also illustrated in Figure 9, is images that do not contain particularly interesting objects. However, here all the saliency maps make errors, which might indicate that some other factors in addition or instead of visual saliency guide eyefixations for this type of images. This explains the gap between the proposed method (AUC of 0.823 and 0.839 on MIT and NUSEF) and the intersubject performance (AUC of 0.894 and 0.883 on MIT and NUSEF), see Table 2.

Overall, the qualitative and quantitative analysis illustrate the high potential of proto-object-based measurements for modeling eye-movements.

Optimal weights for different kind of saliency

In this work, three different saliency measurements are considered: internal contrast-based saliency (Equation 6), external contrast-based saliency (Equa-

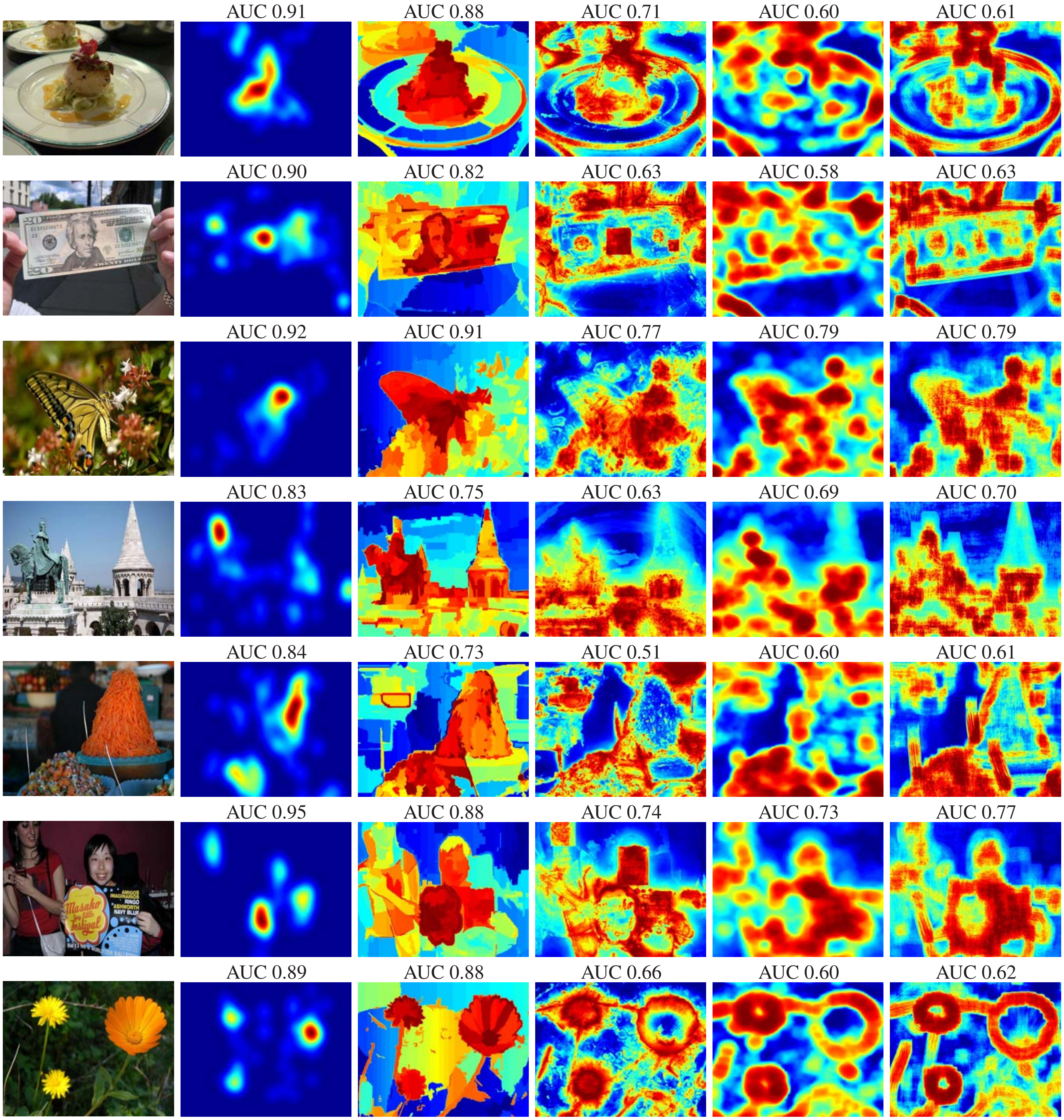


Figure 7. Results for images containing salient objects. From left to right: input image, eye-fixation density map, proto-object-based saliency map (Bruce & Tsotsos, 2009; Hou & Zhang, 2007; Judd et al., 2009).

tion 5), and rarity-based saliency (Equation 9). In the experiments in the Results section, these measurements were combined together linearly with equal weights. However, the results in Table 1 demonstrate that the rarity-based saliency predicts the way people look at images much more accurately than the internal contrast-based saliency (AUC of 0.778 vs. AUC of 0.689 on MIT dataset, and AUC of 0.759 vs. AUC of 0.656 on NUSEF dataset). Therefore, to investigate the optimal way of combining different kinds of saliency, we examine various weighting schemes. Particularly, instead of using Equation 11 to calculate the final saliency of a proto-object, we compute a set saliency scores considering a range of possible weights W_1 and W_2

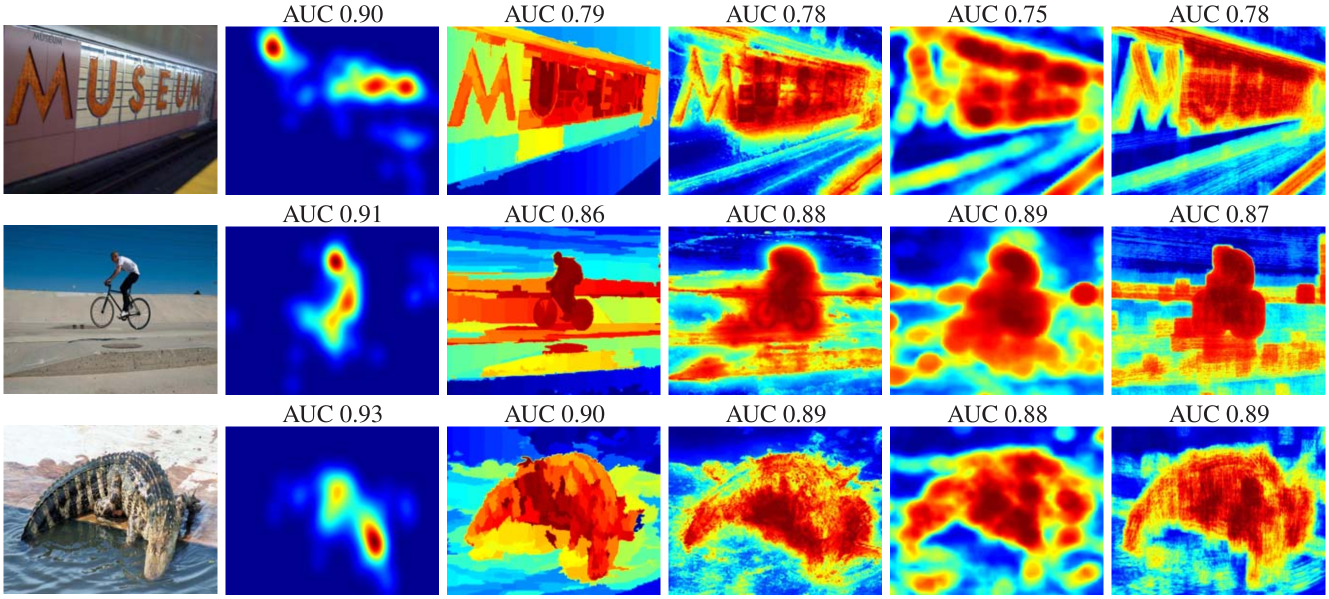


Figure 8. Example of images when all methods have similar performance. From left to right: input image, eye-fixation density map, proto-object-based saliency map (Bruce & Tsotsos, 2009; Hou & Zhang, 2007; Judd et al., 2009).

$$
\begin{aligned}
&Sal_{proto-object}(a, W_1, W_2)\\
&\quad = W_1{*}\Big(W_2{*}Sal_{internal}(a)\\
&\qquad\qquad +(1 - W_2){*}Sal_{external}(a)\Big)\\
&\qquad\qquad +(1 - W_1){*}Sal_{rarity}(a), \qquad (11)
\end{aligned}
$$

where W_1 and W_2 take values from 0 to 1. Therefore, when both W_1 and W_2 are 0, $Sal_{proto-object}(a, W_1, W_2) = Sal_{rarity}(a)$, whereas when both W_1 and W_2 equal to 1, $Sal_{proto-object}(a, W_1, W_2) = Sal_{internal}(a)$.

Figure 10 shows the averaged AUCs for images from the MIT dataset depending on weights W_1 and W_2. The highest AUC of 0.7913 is reached when $W_1 = 0.5$ and $W_2 = 0.2$. This indicates that the external contrast-based and rarity-based saliency have similar impact to the eye-movements allocation, whereas internal contrast-based saliency is less important. However, the highest performance is spread over a range of values. For example, when equal weights are considered, as in Equation 11, the performance is 0.785, and any values of $W_1 = [0.3; 0.6]$ and $W_2 = [0; 0.5]$ are reasonable.

Discussion

The role of an object in visual attention has been explored by many (Einhäuser et al., 2008; Friedman, 1979; Henderson, 2003; Nuthmann & Henderson, 2010; Scholl, 2001; Vincent et al., 2009). In most studies it is assumed that an object is already recognized. Friedman (1979) demonstrated that people focus longer on objects that are out of context. Vincent et al. (2009) advanced the hypothesis that highly visible spots in the image—for example, lantern lights—may attract less eye fixations than less visible, but semantically more informative objects. Nevertheless, it is an important question what happens prior to object recognition. According to the object-based attention theory, an input image is first segmented into proto-objects by feature grouping. Then, the importance of these proto-objects is evaluated, so that an observer looks to the most salient one. As a result, attended proto-object is represented as already recognized object, or an object part. During the recognition step, the initial segmentation may be corrected. In this work, we have concentrated on the proto-object importance at the pre-attentive stage, i.e., before object recognition. We have proposed a way to estimate their bottom-up saliency. The results in Table 2 illustrate that the proto-object-based approach outperforms state-of-the-art computational methods. Therefore, our experiments have confirmed an important role of objects even at the early pre-attentive stage.

Shape and spatial extend of proto-objects

As can be observed in Figures 7–9, the majority of the eye-fixations fall within objects. However, the bigger the object is, the less uniform is the distribution of fixation locations within it. Therefore, instead of highlighting the whole object uniformly, it is essential to estimate saliency in a hierarchical fashion. First, the coarser scale should be used to estimate saliency of the whole object. Then, in order to find salient details

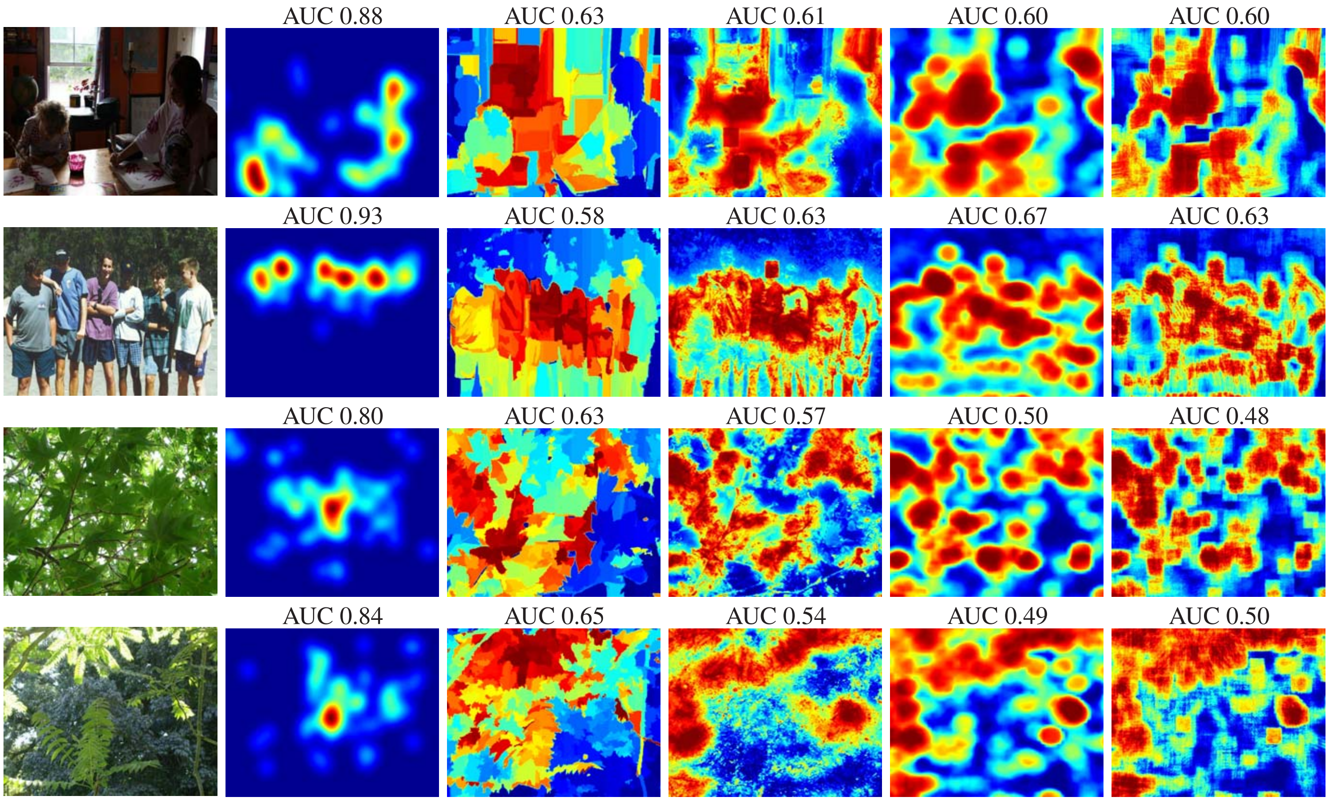


Figure 9. Examples of mistakes. From left to right: input image, eye-fixation density map, proto-object-based saliency map (Bruce & Tsotsos, 2009; Hou & Zhang, 2007; Judd et al., 2009).

within the object, the finer scales should be considered. Furthermore, people rarely fixate on object boundaries. Thus, it might be not so important to identify the exact object borders. We hypothesize that the approximation of proto-objects used in this paper may be sufficient enough to model visual attention.

To investigate the relationship between the eye-movements that fall within proto-objects and the shape and spatial extend of proto-objects, we examine the within-proto-object Preferred Viewing Locations (PVL; Henderson, 1993; Nuthmann & Henderson, 2010; Rayner, 1979). It has been shown that people tend to fixate at the middle of the words while reading (Rayner, 1979), and at the center of the objects while observing line drawings (Henderson, 1993) and natural scenes (Nuthmann & Henderson, 2010). We examine if this behavior can be extended to the proto-objects. Particularly, we analyze the distribution of distances between landing positions and the center of proto-object. Following Nuthmann and Henderson (2010), we consider the horizontal and vertical components separately, which we normalize to the height and width of the bounding box around proto-object. To investigate the influence of the spatial extend of proto-objects, we divide proto-objects in six different groups according to their size. Furthermore, we distinguish proto-objects as more and less salient, where a proto-object is

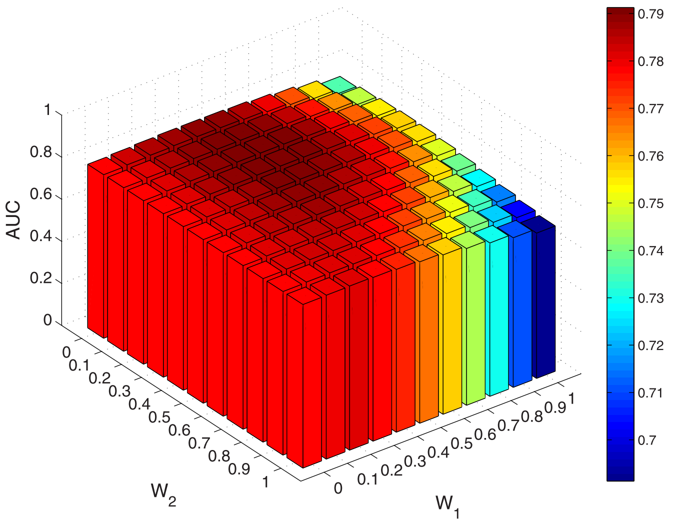
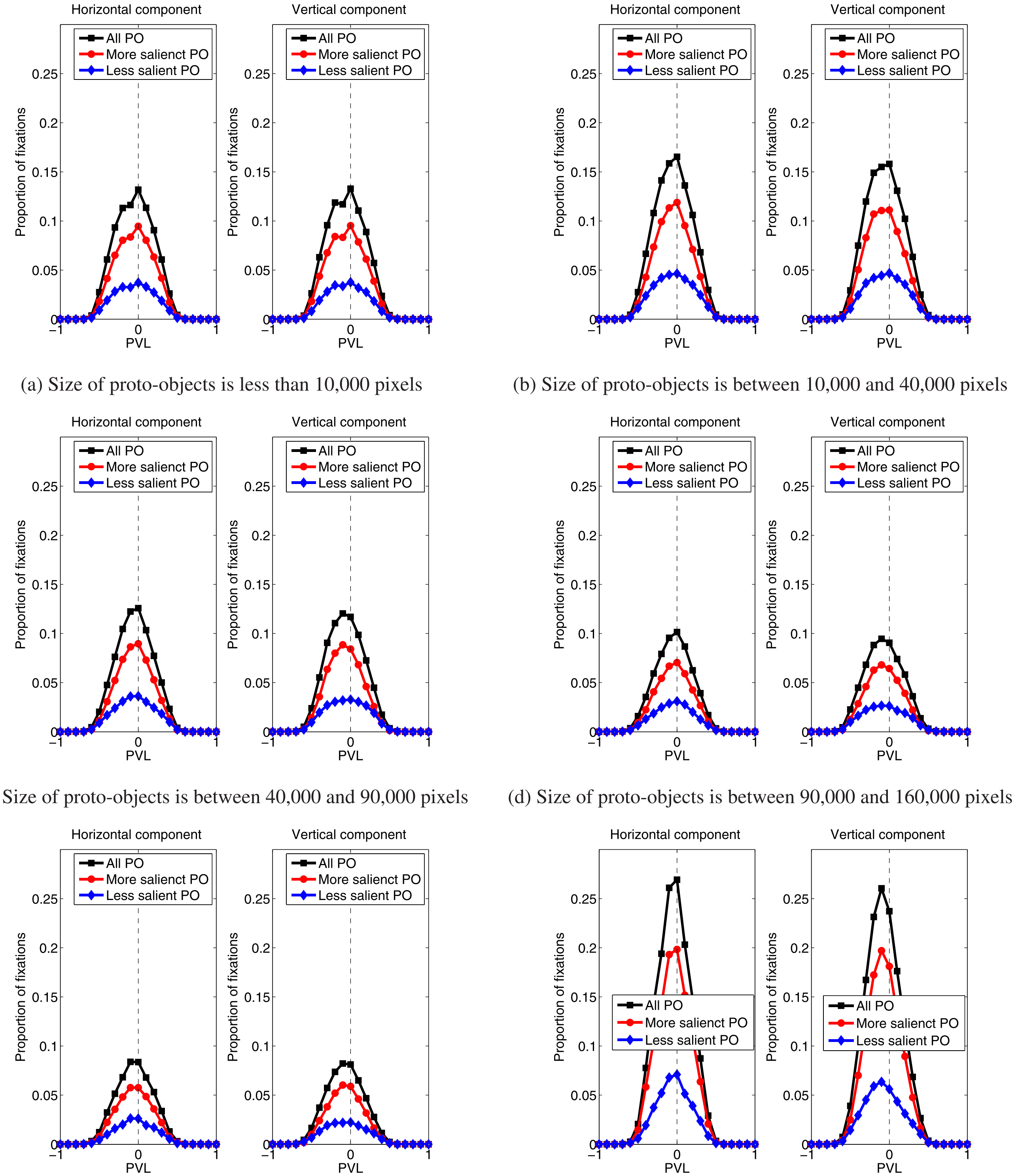


Figure 10. Influence of the different weighting schemes of the contrast-based and rarity-based saliency to the accuracy of the proposed method. W_1 controls the impact of the rarity-based saliency, whereas W_2 defines the contribution of internal and external contrast-based saliency. To achieve high performance, external contrast-based saliency should be weighted higher than internal contrast-based saliency. Furthermore, rarity-based and contrast-based saliency should have nearly equal weights.



(a) Size of proto-objects is less than 10,000 pixels

(b) Size of proto-objects is between 10,000 and 40,000 pixels

(c) Size of proto-objects is between 40,000 and 90,000 pixels

(d) Size of proto-objects is between 90,000 and 160,000 pixels

(e) Size of proto-objects is between 160,000 and 250,000 pixels

(f) Size of proto-objects is more than 250,000 pixels

Figure 11. Preferred viewing locations for proto-objects in images from MIT dataset (Judd et al., 2009). Broken lines correspond to the case when people fixate at the center of the proto-object. The strong evidence for within-proto-objects PVL close to the center of proto-objects confirms the choice of proto-objects as units of saliency analysis.

regarded as more salient if its saliency is above the mean saliency of all proto-objects.

Figure 11 shows the averaged results over all images and fixation locations from MIT dataset (Judd et al., 2009). It can be seen, that in all considered cases the distances between within-proto-object landing positions and the center of proto-objects are normally distributed with a mean around zero. Thus, people tend to look somewhere close to the center of proto-objects throughout all scales.

Conclusion

Our experiments have demonstrated the potential of the proto-objects as units of the analysis. Research in neuroscience (Yantis & Serences, 2003) points out that visual attention may be directed to spatial locations, objects, and even surfaces (Nakayama, He, & Shimojo, 1995). It seems likely that the unit of attention depends on the task, on the field of view, and on the observer's intentions (Scholl, 2001). For example, attention might adopt a spatial-based behavior within complex extended objects, be object-based on the global scale, and be directed to any well-formed perceptually distinguishable surface. Which aspect prevails depends on which of these factors will dominate (Einhäuser et al., 2008). We hypothesize that a complete model for visual attention necessarily incorporates object-based, spatial-based, and surface-based information.

Keywords: visual attention, saliency, eye movements, proto-objects

Acknowledgments

This research was supported by the FP7-STREP DALi project. We thank the anonymous *Journal of Vision* reviewers and Ruth Rosenholtz for their valuable feedback on this work.

Commercial relationships: none.
Corresponding author: Victoria Yanulevskaya.
Email: yanulevskaya@gmail.com.
Address: Department of Information Engineering and Computer Science, University of Trento, Italy.

References

Alexe, B., Deselaers, T., & Ferrari, V. (2010). What is an object? In *IEEE conference on computer vision and pattern recognition* (pp. 73–80).

Baddeley, R. J., & Tatler, B. W. (2006). High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis. *Vision Research*, *46*(18), 2824–2833.

Beck, J. (1966). Effect of orientation and of shape similarity on perceptual grouping. *Attention, Perception, & Psychophysics*, *1*(5), 300–302.

Bruce, N., & Tsotsos, J. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, *9*(3):5, 1–24, http://www.journalofvision.org/content/9/3/5, doi:10.1167/9.3.5. [Pubmed] [Article]

Buswell, G. T. (1935). *How people look at pictures: a study of the psychology and perception in art*. Chicago: University of Chicago Press.

Cerf, M., Frady, E. P., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision*, *9*(12):10, 1–15, http://www.journalofvision.org/content/9/12/10, doi:10.1167/9.12.10. [Pubmed] [Article]

Cheng, M.-M., Zhang, G.-X., Mitra, N. J., Huang, X., & Hu, S.-M. (2011). Global contrast based salient region detection. In *IEEE conference on computer vision and pattern recognition* (pp. 409–416). Colorado Springs, CO: IEEE.

Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision* (Vol. 1, p. 22). Prague: Springer.

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General*, *113*, 501–517.

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review*, *96*(3), 433.

Egly, R., Driver, J., & Rafal, R. D. (1994). Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General*, *123*, 161–177.

Einhäuser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, *8*(2):2, 1–19, http://www.journalofvision.org/content/8/2/2, doi:10.1167/8.2.2. [Pubmed] [Article]

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, *8*(14):18, 1–26, http://www.journalofvision.org/content/8/14/18, doi:10.1167/8.14.18. [Pubmed] [Article]

Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision*, *8*(3):3, 1–15, http://www.journalofvision.org/content/8/3/3, doi:10.1167/8.3.3. [Pubmed] [Article]

Felzenszwalb, P. F., & Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *International Journal of Computer Vision*, *59*(2), 167–181.

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision*, *8*(2):6, 1–17, http://www.journalofvision.org/content/8/2/6, doi:10.1167/8.2.6. [Pubmed] [Article]

Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General*, *108*(3), 316–355.

Gao, D., Mahadevan, V., & Vasconcelos, N. (2008). On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of Vision*, *8*(7):13, 1–18, http://www.journalofvision.org/content/8/7/13, doi:10.1167/8.7.13. [Pubmed] [Article]

Harel, J., Koch, C., & Perona, P. (2007). Graph-based visual saliency. In *Advances in neural information processing systems* (Vol. 19, pp. 545–552). Cambridge, MA: MIT Press.

Henderson, J. M. (1993). Eye movement control during visual object processing: Effects of initial fixation position and semantic constraint. *Canadian Journal of Experimental Psychology*, *47*, 7998.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, *7*(11), 498–504.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., and Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. *Eye movements: A window on mind and brain* (pp. 537–562). Amsterdam: Elsevier.

Horowitz, T. S., Fine, E. M., Fencsik, D. E., Yurgenson, S., & Wolfe, J. M. (2007). Fixational eye movements are not an index of covert attention. *Psychological Science*, *18*(4), 356.

Hou, X., & Zhang, L. (2007). Saliency detection: A spectral residual approach. In *IEEE conference on computer vision and pattern recognition* (pp. 1–8). Minneapolis, MN: IEEE.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *International Conference on Computer Vision* (pp. 2106–2113). Kyoto, Japan: IEEE.

Julesz, B. (1981). Textons, the elements of texture perception, and their interactions. *Nature*, *290*, 91–97.

Julesz, B., & Bergen, J. R. (1983). Textons, the fundamental elements in preattentive vision and perception of textures. *Bell System Technical Journal*, *62*, 1619–1645.

Kadir, T., & Brady, M. (2001). Saliency, scale and image description. *International Journal of Computer Vision*, *45*(2), 83–105.

Kelley, T. A., Serences, J. T., Giesbrecht, B., & Yantis, S. (2008). Cortical mechanisms for shifting and holding visuospatial attention. *Cerebral Cortex*, *18*(1), 114.

Kienzle, W., Franz, M. O., Scholkopf, B., & Wichmann, F. A. (2009). Center-surround patterns emerge as optimal predictors for human saccade targets. *Journal of Vision*, *9*(5):7, 1–15, http://www.journalofvision.org/content/9/5/7, doi:10.1167/9.5.7. [Pubmed] [Article]

Krieger, G., Rentschler, I., Hauske, G., Schill, K., & Zetzsche, C. (2000). Object and scene analysis by saccadic eye-movements: An investigation with higher-order statistics. *Spatial Vision*, *2*(3), 201–214.

Lang, P. J., Bradley, M. M., and Cuthbert, B. N. (1999). *International affective picture system (IAPS): Technical manual and affective ratings*. Gainesville, FL: The Center for Research in Psychophysiology, University of Florida.

Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., et al. (2010). Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(2), 353–367.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1996). The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision*, *10*(3), 165–188.

Nakayama, K., He, Z. J., & Shimojo, S. (1995). Visual surface representation: A critical link between lower-level and higher-level vision. *Visual Cognition: An Invitation to Cognitive Science*, *2*, 1–70.

Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal* *of Vision*, *10*(8):20, 1–19, http://www.journalofvision.org/content/10/8/20, doi:10.1167/10.8.20. [Pubmed] [Article]

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175.

Orabona, F., Metta, G., & Giulio, S. (2007). A proto-object based visual attention model. *Attention in Cognitive System*, *4840*, 198–215.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, *42*(1), 107–123.

Posner, M. I. (1980). Orienting attention. *Quarterly Journal of Experimental Psychology*, (32), 3–25.

Rayner, K. (1979). Eye guidance in reading: Fixation locations within words. *Perception*, *8*(1), 21.

Reinagel, P., & Zador, A. (1999). Natural scene statistics at the centre of gaze. *Network: Computation in Neural Systems*, *10*(4), 341–350.

Renninger, L. W., Coughlan, J., Verghese, P., & Malik, J. (2005). An information maximization model of eye movements. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems* (Vol. 17, pp. 1121–1128). Cambridge, MA: MIT Press.

Renninger, L. W., Verghese, P., & Coughlan, J. (2007). Where to look next? Eye movements reduce local uncertainty. *Journal of Vision*, *7*(3):6, 1–17, http://www.journalofvision.org/content/7/3/6, doi:10.1167/7.3.6. [Pubmed] [Article]

Rensink, R. (2000). Seeing, sensing, and scrutinizing. *Vision Research*, *40*(10–12), 1469–1487.

Rosenholtz, R. (1999). A simple saliency model predicts a number of motion popout phenomena. *Vision Research*, *39*(19), 3157–3163.

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision*, *77*(1), 157–173.

Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition*, *80*(1–2), 1–16.

Simoncelli, E. P., & Freeman, W. T. (1995). The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *International Conference on Image Processing* (Vol. 3, pp. 444–447). Washington, DC: IEEE.

Sivic, J., & Zisserman, A. (2003). Video Google: A text retrieval approach to object matching in videos. In *International Conference on Computer Vision* (pp. 1470–1477). Nice, France: IEEE.

Subramanian, R., Katti, H., Sebe, N., Kankanhalli, M., & Chua, T. S. (2010). An eye-fixation database for saliency detection in images. *European Conference on Computer Vision*, IV, 30–43.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, *7*(14):4, 1–17, http://www.journalofvision.org/content/7/14/4, doi:10.1167/7.14.4. [Pubmed] [Article]

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, *45*(5), 643–659.

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, *11*(5):5, 1–23, http://www.journalofvision.org/content/11/5/5, doi:10.1167/11.5.5. [Pubmed] [Article]

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Uijlings, J. R. R., Smeulders, A. W. M., & Scha, R. J. H. (2009). Real-time bag of words, approximately. In *International Conference on Image and Video Retrieval* (pp. 1–8). Santorini Island, Greece: IEEE.

Underwood, G., Foulsham, T., Van Loon, E., Humphreys, L., & Bloyce, J. (2006). Eye movements during scene inspection: A test of the saliency map hypothesis. *European Journal of Cognitive Psychology*, *18*(3), 321–342.

van de Sande, K. E. A., Uijlings, J. R. R., Gevers, T., & Smeulders, A. W. M. (2011). Segmentation as selective search for object recognition. In *International Conference on Computer Vision* (pp. 1879–1886). Barcelona, Spain: IEEE.

Vecera, S. R., & Farah, M. J. (1994). Does visual attention select objects or locations? *Journal of Experimental Psychology: General*, *123*, 146–160.

Vincent, B., Baddeley, R., Correani, A., Troscianko, T., & Leonards, U. (2009). Do we look at lights? Using mixture modeling to distinguish between low-and high-level factors in natural image viewing. *Visual Cognition*, *6*(7), 856–879.

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks*, *19*(9), 1395–1407.

Wischnewski, M., Belardinelli, A., Schneider, W. X., & Steil, J. J. (2010). Where to look next? combining static and dynamic proto-objects in a TVA-based model of visual attention. *Cognitive Computation*, *2*(4), 326–343.

Yantis, S., & Serences, J. T. (2003). Cortical mechanisms of space-based and object-based attentional control. *Current Opinion in Neurobiology*, *13*(2), 187–193.

Yanulevskaya, V., Uijlings, J. R. R., & Geusebroek, J. M. (2013). Salient object detection: From pixels to segments. *Image and Vision Computing*, *31*(1), 31–42.

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, *8*(7):32, 1–20, http://www.journalofvision.org/content/8/7/32, doi:10.1167/8.7.32. [Pubmed] [Article]

Zhao, Q., & Koch, C. (2011). Learning a saliency map using fixated locations in natural scenes. *Journal of Vision*, *11*(3):9, 1–15, http://www.journalofvision.org/content/11/3/9, doi:10.1167/11.3.9. [Pubmed] [Article]