Modelling human visual attention: from pixels to proto-objects
Yanulevskaya, V.

Citation for published version (APA):
Yanulevskaya, V. O. (2012). Modelling human visual attention: from pixels to proto-objects

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

Download date: 02 Jan 2019
PROTO-OBJECT-BASED COMPUTATIONAL MODEL FOR VISUAL ATTENTION

State-of-the-art bottom-up saliency models are based on the spatial-based attention theory [18], where a pixel or patch of an image serves as the unit of analysis. However, people do not perceive an image as a collection of pixels but rather as a collection of entities. According to the object-based theory [64], brain groups similar pixels into coherent regions already at the pre-attentive stage when objects are not yet recognized. These regions are called proto-objects. While looking around, people shift their attention from one proto-object to another. Thus, instead of evaluating saliency at the pixel level, we propose a computational model that measures saliency at the proto-object level. As proto-objects are extracted in the pre-attentive stage, they do not necessarily correspond to real objects precisely, they also can outline only a distinct part of an object, or a group of similar objects. Therefore, proto-objects are hierarchically ordered and can occur at all possible scales in the image. We take this into account by using a hierarchical image segmentation to generate a set of proto-objects. Afterwards, the saliency of these proto-objects is estimated and combined into a proto-object-based saliency map. We evaluate the proposed method on two challenging eye-fixation datasets [38] and [79]. The results demonstrate that the proto-object-based approach outperforms state-of-the-art spatial-based models on predicting human fixations.

5.1 INTRODUCTION

To understand the processes that control eye-fixations it is important to determine what causes the shifts of focus of attention. Do people attend to spatial locations or to discrete objects while looking around?

Traditionally, visual attention is compared to a spotlight or a zoom lens. Every time our focus is moved to a new location the information within a fovea-sized region around it is processed [18, 59, 84]. Such a
region may contain any mixture of objects and background. Therefore, within the spatial-based theory attention shifts from one spatial location to another, where the shape of the attended regions is not influenced by the visual content. Alternatively, within the object-based attention theory [64] the visual structure of each object plays a vital role in the way the attention spreads out. The theory assumes that an attended area coincides with a proto-object, which is defined as a coherent region which approximates an object, a part of object, or group of objects [70, 91]. In this chapter, we follow the latter theory to propose a novel proto-object-based computational model for visual attention.

![Image of input image, human fixations, and proposed saliency map.](image)

**Figure 5.1:** While people tend to spread their attention within an object (b), spatial-based computational models for visual attention generally highlight only parts of an object with high contrast (d)-(f). We propose to overcome it 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.

The difficult part of building an object-based computational model is to separate objects from the background. The methods which require precise object locations [15, 54, 90] are constrained by manual object outlining because an accurate automatic segmentation for generic objects requires so much a priori knowledge that it is still beyond current techniques. Moreover, an accurate automatic segmentation implies that all objects are already recognized, whereas attention is believed to start acting
before object recognition [91]. To relax this requirement for precise object locations, we consider proto-objects as units of analysis. According to the object-based attention theory, the visual system pre-segments a scene into coherent regions by feature grouping at the early pre-attentive stage. Recent research in object detection [88] and results from Chapter 4 of this thesis have shown that approximations of object locations can be successfully used in practical applications. Therefore, we adapt the hierarchical image segmentation used in [88] to extract proto-objects of an image.

Based on the intuition that visually salient objects attract more attention than non-salient objects, the importance of a proto-object can be measured by estimating how much the proto-object "pops out" from the image. From the literature, it is known that an image region pops out in two cases: (1) when it differs from the surroundings [12,36,41,42,45,61], and (2) when it contains rare or outstanding details [2,6,23,95]. In this work, both contrast-based and rarity-based saliency are considered. Moreover, the notion of an object is directly incorporated into saliency measurements as proto-objects are used as units of analysis. Thereby, it allows for highlighting important image regions in their entirety, whereas, as is illustrated in Figure 5.1, measurements at the pixel level mostly highlight specific details.

In this chapter we are building on the approach for salient object detection proposed in Chapter 4. In the previous chapter, the aim is to select the most salient proto-object that captures a complete object. In the current work, we investigate the link between saliency of proto-objects and the way people look at images. Particularly, the hypothesis is that people shift their attention from one proto-object to another based on their saliency. Therefore, the more salient a proto-object is, the more fixations it will attract. Thus in this chapter, 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 Chapter 4, we extend the measurement of the contrast-based saliency by introducing external and internal contrast of a proto-object.

Our main research questions are: (1) Does saliency of proto-objects predict the way people look at images? (2) Does a proto-object-based approach predict human fixations better than spatial-based approaches? We investigate these questions using two standard eye-fixation datasets [38] and [79].
5.2 RELATED WORK

Though some object-based models for visual attention have been proposed, the biggest effort has been made to investigate spatial-based attention \[6,10,23,30,36,39,41,62,65,81,97\]. Thus, all such models construct a saliency map of an entire image by measuring the importance of individual pixels or patches. Itti et al. \[36\] 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 colour, 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. Another view on spatial-based saliency is introduced by Bruce and Tsotsos \[6\]. They define saliency based on the information maximization principle. Intuitively, image locations with unexpected content in comparison with their surrounding are more informative, and thus salient. Hou and Zhang \[33\] reversed the problem and considered the minimization of the energy consumption in the brain as the main goal of visual attention. We will demonstrate that the proposed proto-object-based method consistently outperforms these spatial-based approaches in saliency prediction.

Einhauser et al. \[15\] investigate the role of objects in visual attention. In their experiments they manually segmented images to localize objects. Then, in addition to eye-fixation recording, they asked subjects to name the most interesting objects within an image. Weighted with recall frequency, locations of these objects predict eye-fixations better than the standard spatial-based method by Itti et al. \[36\]. We move one step further by replacing manual segmentation with automatical hierarchical grouping to extract proto-objects from the image.

The correlation between the significance of an object and its saliency has been demonstrated by Elazary and Itti \[16\]. In their experiments, Elazary and Itti \[16\] considered the LabelMe database \[69\] 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, segmented objects are considered to be interesting. The authors demonstrated that the high peaks of the saliency map \[36\] coincide with the segmented objects. This implies that objects which attract attention tend to contain visually outstanding details. Therefore we measure how much proto-objects pop out from the scene to predict where people look.
In contrast to \cite{15,16}, Walther and Koch \cite{91} proposed a way to use spatial-based saliency map for generating proto-objects automatically. They considered the pixel level saliency map of \cite{36} and determined the spatial extent of its peaks as proto-objects. Particularly, an extracted proto-object is a set of neighbouring 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 objects is not taken into account explicitly. In this chapter, the proto-objects are extracted directly from the image by hierarchical segmentation. Therefore, the structure of objects of an image determines the shape of proto-objects.

In this chapter we make the following contributions: (1) 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. (2) To extract proto-objects we segment the image, rather than its derivative pixel level saliency map. It allows for the creation of a proto-object level saliency map as a combination of the saliency of all proto-objects. (3) We demonstrate that the proposed method outperforms the state-of-the-art in saliency prediction on two challenging eye-fixation datasets.

5.3 Saliency map based on proto-objects

A proto-object is a coherent image region which, by the visual coherence in most objects in the world, will roughly corresponds to part of an object, a complete object, or a group of objects. Hence, an object of an image may consist of several small proto-objects approximating its parts, and, at the same time, be part of a larger proto-object which 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 Chapter 4, we adapt the hierarchical image segmentation used in \cite{88} to extract proto-objects. Afterwards the saliency of all segments is combined into the final saliency map.

5.3.1 Proto-objects extraction

For the proto-object extraction, we use the same approach discussed in Section 4.3.1 of Chapter 4. However, in Chapter 4 the task of salient object detection is considered. Therefore, it is beneficial to combine multiple hierarchical segmentations, as it increases chances to generate a segment accurately outlining the object. In contrast, this chapter investigates the way people look at images. Nuthmann and Henderson \cite{54} have shown
that one’s preferred viewing locations are close to the center of objects. Since it is not so important to identify the exact borders of objects, the use of only one hierarchical segmentation is sufficient. Figure 5.2 provides two examples of the hierarchy of proto-objects.

Figure 5.2: 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 neighbouring segments are merged according to their similarity (see (c) and (k)), we obtain proto-objects which correspond to the objects and small groups of similar objects: for example buildings and districts in (g), or individual persons and groups of people which 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.
5.3.2 Proto-object saliency estimation

To measure the saliency of a proto-object we estimate how much it pops out off the image. An image region pops out when it differs from the surroundings and when it contains rare or outstanding details. 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 their saliency and at the same time encourage proto-objects which approximate image objects more accurately. Recently, Cheng et al. have proposed a global contrast-based method to estimate saliency of a region. They segment an image into the set of non-intersecting regions. Then they estimate saliency of a region by measuring the contrast between this region and all other regions in the image. In this chapter, we extend their approach to hierarchically overlapping regions.

Let $P$ be an initial image segmentation. We calculate the saliency of a proto-object $a$ as a sum of colour 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 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 neighbouring proto-objects is more important than between remote ones. Therefore, the contrast is weighted with the spatial distance between proto-objects. Furthermore, the contrast to larger segments influences proto-object saliency more than the contrast to smaller segments. Therefore, the proto-object saliency based on the external contrast is defined as:

$$\text{Sal}_{\text{external}}(a) = \sum_{p_i \in P \setminus a} C(a, p_i) \exp\left(-\frac{D(a, p_i)}{\sigma^2} \right) \frac{|p_i|}{|P| - |a|},$$

where $p_i$ are segments from the initial segmentation $P$, $C(a, p_i)$ is the contrast between proto-objects $a$ and $p_i$, which is weighted with the Euclidean distance between the centroids of the corresponding proto-objects $D(a, p_i)$ and with the number of pixels $|p_i|$ within the segment $p_i$. The parameter $\sigma$ controls the contribution of remote segments to saliency estimation. The number of pixels outside $a$ $|P| - |a|$ is a normalization factor.
Furthermore, the saliency of a proto-object depends on the complexity of the proto-object itself\textsuperscript{[39, 63]}. A piece of sky may differ considerably from the rest of an image and still not pop out because of its uniform structure. Therefore, to estimate the complexity of a proto-object, its parts are compared with one another. Particularly, we again consider the initial image segmentation $P$ to estimate the average difference between all segments within a proto-object. We call it \textit{internal contrast}. 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:

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

(5.2)

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 contrast into the contrast-based saliency:

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

(5.3)

To create a contrast-based saliency map we average the saliency of proto-objects calculated according to Eq. (5.3) over the pixels they cover. Figure 5.3 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.

\textit{Rarity-based saliency}

Rarely-occurring image structures attract attention\textsuperscript{[6, 23]}. Therefore they should be highlighted. In contrast, frequently-occurring details are typically part of the image background left unattended. They should be suppressed. To capture image structures we represent an image as a bag-of-visual-words\textsuperscript{[14, 76]}, which is the state-of-the-art technique in object detection. In this representation an image is divided into patches using a regular grid. Then each patch is represented by the SIFT descriptor\textsuperscript{[46]}, which efficiently captures both contours and texture of an image patch by summation of its gradient orientations. The set of SIFT descriptors is quantized where the clusters are called \textit{visual words}. Thereby, each patch of an image is mapped to a visual word so that an image may be represented as a \textit{bag-of-visual-words}\textsuperscript{[14, 76]}. As an image representation
5.3 Saliency Map Based on Proto-Objects

<table>
<thead>
<tr>
<th>Input image</th>
<th>Eye fixations</th>
<th>External contrast SM</th>
<th>Internal contrast SM</th>
<th>Combined contrast SM</th>
</tr>
</thead>
</table>

**Figure 5.3:** 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 colour, 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.

In terms of SIFT visual words is currently the most effective one for the recognition of objects, such 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, we follow the information maximization approach [6] calculating the saliency of a pixel $p_i$ as the self-information of the corresponding visual word $w_{p_i}$:

$$\text{Sal}_{\text{pixel}}(p_i) = -\log(P(w_{p_i})),$$

where $P(w_{p_i})$ is the probability of a visual word $w_{p_i}$.

Eq. (5.4) defines saliency at the pixel level. To incorporate the notion of object into the measurements, we define rarity-based saliency at the proto-object level by averaging the saliency of all its encompassing pixels:

$$\text{Sal}_{\text{rarity}}(a) = \frac{1}{|a|} \sum_{p_i \in a} \text{Sal}_{\text{pixel}}(p_i),$$

where $|a|$ is the number of pixels within a proto-object $a$. As in the previous section 5.3.2 we create a rarity based proto-object saliency map by averaging the saliency of proto-objects over the pixels they cover. Figure 5.4 illustrates the difference between the rarity-based saliency at pixel and proto-object levels. While pixel-based measurements highlight mostly object edges, Eq. (5.5) smoothes the saliency over proto-objects, thereby effectively highlighting entire objects.
Figure 5.4: An example 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.

**Combined proto-object-based saliency map**

In this section saliency measurements are combined into the final saliency map. Contrast-based saliency measures how much proto-objects pop out from the surrounding based on colour information. Rarity-based saliency measures when there are unique details within proto-objects based on texture information. These two measurements are complementary to one another. Therefore, we combine contrast-based and rarity-based saliency to the final saliency measurement:

$$\text{Sal}_{\text{proto-object}}(a) = \text{Sal}_{\text{contrast}}(a) + \text{Sal}_{\text{rarity}}(a).$$  \hspace{1cm} (5.6)

This way we estimate saliency for all proto-objects. Then we average proto-object saliency over the pixels they cover, because proto-objects have a hierarchical structure and may be overlapping. The resulting saliency map predicts where people focus their attention while investigating an image.

5.3.3 **Implementational details**

**Proto-object extraction.** In the same way as in Chapter 4, we run Felzenswalb and Huttenlocher’s algorithm [19] to over-segment an image. However, in this Chapter we segment the image only once with the
5.4 Evaluation

We test the proto-object-based method on two recent eye-fixation datasets: MIT \[38\] and NUSEF \[79\], 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 \[32,40\]. However in everyday viewing conditions, they can be tightly linked \[58\].

5.4.1 Evaluation method

The standard evaluation in eye-fixation prediction \[6,38,81\] 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, for each binary map the true positive and false positive rates are calculated. The true positive rate is the fraction of fixated pixels above threshold, while the false positive rate is the fraction of non-fixated pixels above 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 accuracy of the prediction. For the perfect saliency map the AUC is 1, and for a random saliency map the AUC is 0.5. However, given the variability in the way people look at the same image, it is impossible to achieve an AUC of 1. Rather, we consider the inter-subject predictive power as the upper-bound. Similarly to [38, 97], we estimate how well the fixations of one subject can be predicted by the fixations of the rest of the subjects convolved with a fovea-sized two-dimensional Gaussian kernel ($\sigma = 1^\circ$, i.e. 30 pixels). To calculate the performance of a computational model we estimate how well the fixations of one subject are predicted by the generated saliency map. We calculate AUCs for all subjects and images in the dataset. The average results are reported. To allow for statistical inference, we calculate the Standard Error (SE) of the AUC based on the method proposed in [29]. In addition, we use a two-tailed $t$-test to report the significance of the difference in performance of various model predictions.

5.4.2 The eye-fixation datasets

We consider two eye-fixation datasets: MIT [38] and NUSEF [79], 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. [38] contains 1003 images and eye-fixations of 15 subjects. All images are randomly selected from Flickr and LabelMe [69] and have diverse appearances of everyday scenes: they range from landscapes and portraits to close-ups and graffiti. For this reason the dataset is representative and challenging. Figure 5.5 illustrates a number of images from the MIT dataset.

The NUSEF dataset by Subramanian et al. [79] contains 751 images and fixations of on average 24 subjects. 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 [18] dataset. Figure 5.6 shows representative images from the NUSEF dataset.
5.4 Evaluation

Figure 5.5: Example of images from MIT dataset [38].

Figure 5.6: Example of images from NUSEF dataset [79].

5.4.3 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 spatial-based approaches of [6, 33, 36] and with the inter-subject variability which sets the upper-boundary of the performance. All results are presented in Table 5.1.

The internal contrast-based saliency alone achieves a moderate AUC of 0.689 on MIT dataset and 0.656 on NUSEF dataset, see Table 5.1. This type of saliency measures the difference between the proto-object itself and its parts. This allows 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 external contrast-based saliency, which estimates how much the proto-object varies from its surrounding. As it is shown in Table 5.1, 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. Thereby, the proto-object contrast-based saliency already outperforms the spatial-based approaches, the t-test indicates that AUCs for the combined contrast-based saliency map are statistically significantly higher that AUCs for spatial-based saliency maps \([6,33,36]\) with \(p \leq 0.0037\).

**Table 5.1:** ROC areas of different methods compared in this chapter, standard errors calculated according to \([29]\) are given in parentheses. The internal contrast-based saliency does not perform as good as state-of-the-art spacial-based methods \([36,6,33]\), but combined contrast-based saliency and especially rarity-based saliency at proto-object level achieve good performance. The combined proto-object-based saliency has the best results.

<table>
<thead>
<tr>
<th>Type of Saliency Map</th>
<th>MIT ([38])</th>
<th>NUSEF ([79])</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal contrast-based sal.</td>
<td>0.689(0.0054)</td>
<td>0.656(0.0051)</td>
</tr>
<tr>
<td>External contrast-based sal.</td>
<td>0.736(0.0048)</td>
<td>0.734(0.0042)</td>
</tr>
<tr>
<td>Combined contrast-based sal.</td>
<td>0.748(0.0047)</td>
<td>0.746(0.0043)</td>
</tr>
<tr>
<td>Rarity-based sal. at pixel level</td>
<td>0.744(0.0047)</td>
<td>0.713(0.0045)</td>
</tr>
<tr>
<td>Rarity-based sal. at proto-object level</td>
<td>0.778(0.0043)</td>
<td>0.759(0.0040)</td>
</tr>
<tr>
<td>Combined proto-object-based sal.</td>
<td>0.785(0.0042)</td>
<td>0.770(0.0039)</td>
</tr>
<tr>
<td>Itti <em>et al.</em> sal. ([36])</td>
<td>0.716(0.0051)</td>
<td>0.686(0.0040)</td>
</tr>
<tr>
<td>Bruce and Tsotsos sal. ([6])</td>
<td>0.735(0.0048)</td>
<td>0.716(0.0048)</td>
</tr>
<tr>
<td>Hou and Zhang sal. ([33])</td>
<td>0.742(0.0050)</td>
<td>0.723(0.0047)</td>
</tr>
<tr>
<td>Inter-subject variability</td>
<td>0.894(0.0023)</td>
<td>0.883(0.0022)</td>
</tr>
</tbody>
</table>

When we compare the rarity-based saliency at the pixel and proto-object level, the differences are even more pronounced. Table 5.1 illustrates that the results of the rarity-based saliency at the pixel level is comparable with the method by Bruce and Tsotsos \([6]\), the null hypothesis of the t-test is rejected with \(p = 0.1337\) and \(p = 0.3027\) on MIT and NUSEF respectively. This is not surprising as both approaches are based on information maximization theory. However, when the saliency is effectively spread over proto-objects, the performance significantly rises from 0.744 to 0.778 \(p < 10^{25}\) on MIT and from 0.713 to 0.759 \(p < 10^{38}\)
5.4 Evaluation

This illustrates the power of proto-objects as units of analysis.

As can be seen in Table 5.1, the combination of contrast- and rarity-based saliency achieves AUC of 0.785 and 0.770 on MIT and NUSEF respectively, which is significantly higher than performance of all individual components of the method ($p \leq 0.00159$). Therefore, contrast-based and rarity-based measurements are complimentary and both are important for saliency prediction. Our final result significantly outperforms the spatial-based approaches on both MIT and NUSEF datasets. Figures 5.7-5.9 provide visual comparisons of the proto-object-based saliency map with the best spatial-based saliency map by Hou and Zhang [33] and with human eye-fixations. Not surprisingly, the advantage is most pronounced for images containing interesting objects. Figure 5.7 demonstrates that although some outstanding features might make an object interesting, people do not only fixate on the most salient details of the object. They tend to inspect the object more thoroughly. For example, in Figure 5.7, the most distinguished detail of a dish, shown in the first row, is a flower on top of it, as is correctly captured by both saliency detectors. Nevertheless people, possibly attracted by the flower, examine the less outstanding parts of the dish as well, expanding saliency of object details to the whole object. The proto-object-based method succeeds in mimicking this behaviour.

In some cases when an image contains a wireframe or a textured object on uniform background, the spatial-based method manages to highlight the whole object as well, see Figure 5.8. Then, the performance of the spatial-based method increases reaching the level of the proto-object-based method.

However, we do not claim that the proposed approach explains eye-movements in all possible situations. Attention is strongly affected by cognitive factors. Therefore modeling of top-down factors is necessary for a complete understanding of the way attention works. In this work we concentrate on bottom-up saliency. Thus our method is not designed for images with objects that are semantically attractive rather than visually. Some examples of such images are given in Figure 5.9. Another difficulty is illustrated in Figure 5.9 where images do not contain particularly interesting objects. In this case, the spatial-based approach makes errors as well, which might indicate that some other factors in addition or instead of visual saliency guide attention for this type of images. This explains the gap between the proposed method (AUC of 0.785 and 0.770 on MIT and NUSEF) and the inter-subject performance (AUC of 0.894 and 0.883 on MIT and NUSEF). However, the main focus of the chapter is
to show that object-based measurements are more suitable for modeling attention than pixel-based ones.

As discussed in Chapter 4, the central part of an image attracts attention \[6, 38, 80\]. Moreover, the Gaussian blob centered in the middle of the image usually shows excellent results which outperforms automatic models of attention \[38, 80, 97\]. Thus, we estimate the influence of the central bias to the evaluated methods. The results in Table 5.2 clearly show that the performance of all models is significantly improved with center bias modeling. However, the ranking of the performances of the models stays the same as shown in Table 5.1, the proto-objects-based saliency achieves the best performance on both datasets (AUC of 0.823 and 0.839 on MIT and NUSEF respectively).

**Table 5.2:** Results of considered saliency models when combined the the central bias. Although adding the central bias boosts the performance of all models, the results of the proto-object-based model are considerably better than the results of spatial-based models.

<table>
<thead>
<tr>
<th>Type of Saliency Map</th>
<th>MIT [38]</th>
<th>NUSEF [79]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central bias alone</td>
<td>0.795 (0.0040)</td>
<td>0.808 (0.0034)</td>
</tr>
<tr>
<td>Proto-object-based saliency with CB</td>
<td>0.823 (0.0037)</td>
<td>0.839 (0.0035)</td>
</tr>
<tr>
<td>Itti et al. saliency [36] with CB</td>
<td>0.803 (0.0047)</td>
<td>0.815 (0.0045)</td>
</tr>
<tr>
<td>Bruce and Tsotsos saliency [6] with CB</td>
<td>0.810 (0.0040)</td>
<td>0.823 (0.0041)</td>
</tr>
<tr>
<td>Hou and Zhang saliency [33] with CB</td>
<td>0.814 (0.0044)</td>
<td>0.820 (0.0041)</td>
</tr>
</tbody>
</table>

5.5 DISCUSSION AND CONCLUSIONS

The role of an object in visual attention has been explored by many \[15, 22, 31, 54, 70, 90\]. In most studies it is assumed that an object is already recognized. Friedman \[22\] demonstrated that people focus longer on objects which are out of context. Vincent et al. \[90\] advanced the hypothesis that highly visible spots in the image, for example lantern lights, may attract less attention 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 visual attention is directed to the most salient one. As a result, attended proto-
objects are represented at a higher level as already recognized objects, or object parts. During the recognition step, the initial segmentation may be corrected. In this chapter, 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 5.1 illustrate that the proto-object-based approach outperforms significantly a spatial-based approaches. Therefore, our experiments have confirmed an important role of objects even at the early pre-attentive stage.

As can be observed in Figures 5.7-5.9, the majority of the eye-fixations fall within objects. However, the bigger the object we consider, the less uniform is the distribution of fixations within it. Therefore, instead of highlighting the whole object uniformly, it is essential to estimate saliency in a hierarchical fashion to place accents within salient objects. Furthermore, people rarely fixate on object contours directing their attention inside interesting areas. Thus, it might be not so important to identify the exact object borders. Therefore, we hypothesize that the rough approximations used in this chapter may work sufficient enough.

Our experiments have demonstrated the advantage of the proto-object-based approach in comparison with the spatial-based approach. Research in neuroscience [92] points out that these two approaches may not be mutually exclusive. Visual attention may be directed to spatial locations and objects. Scholl [70] speculated some time ago that attention demonstrates a spatial-based behavior within complex extended objects [53]. It seems likely that the unit of attention depends on the task, on the field of view, and on the observer’s intentions [70]. Which aspect prevails depends on which of these factors will dominate [15]. We hypothesize that a complete model for visual attention necessarily incorporates both object-based and spatial-based information.
Figure 5.7: Results for images containing salient objects. From left to right: input image, eye-fixation density map, proto-object-based saliency map, saliency map by [33]. Whereas the spatial-based method makes mistakes and highlights salient objects only partially, the proto-object-based method predicts the distribution of fixations more accurate.
Figure 5.8: Example of images when the spatial-based method performs similar to the proto-object-based method. From left to right: input image, eye-fixation density map, proto-object-based saliency map, Hou and Zhang’s saliency map [33].
Figure 5.9: Examples of mistakes. From left to right: input image, eye-fixation density map, proto-object-based saliency map, Hou and Zhang’s saliency map \cite{33}. 