Abstract

The problem of finding locations where people look at first in images, known as saliency detection, spans decades of research from multiple disciplines including psychology, neuroscience, and computer vision. Because of the complexity of the problem it can hardly be considered as solved. Here, we give an overview of the methods for saliency detection starting from early biologically-plausible methods from neuroscience and ending with supervised methods using machine-learning techniques in computer vision. We divide saliency detection methods into three main categories: psychologically-inspired, segmentation-based, and data-driven. We review representative works in each category. Also, we discuss challenges such as evaluation criteria, repeatability and robustness of the methods, and list a number of possible directions for future work.
## Contents

1. Introduction .................................................. 3
2. Physiologically-inspired saliency detection .......................... 4
   2.1 Feature-integration theory ................................ 4
   2.2 Winner-take-all ........................................... 5
   2.3 Information maximization ................................. 8
   2.4 Spectral methods ........................................ 9
   2.5 Context as a saliency cue ............................... 10
3. Segmentation-based methods .................................. 10
   3.1 Global contrast saliency methods ......................... 12
   3.2 Saliency by composition ................................ 12
   3.3 Multi-cue contrast methods for object proposals .... 13
   3.4 Graph-based objects proposals with ranking .......... 14
   3.5 Hierarchical segmentation proposals with ranking .... 15
4. Data-driven methods ........................................ 15
   4.1 SVMs for saliency detection ............................. 16
   4.2 Low-rank factorization ................................ 17
   4.3 Deep learning ........................................... 17
   4.4 Other approaches ....................................... 18
5. Challenges and future work .................................. 19
   5.1 Challenges and open issues ............................ 19
   5.2 Future work ............................................. 20
6. Conclusions ................................................... 21
1 Introduction

The human nervous system's remarkable ability to process visual information extremely fast has always puzzled scientists. Such an ability was necessary for evolutionary purposes as seeing prey or predators fast allowed humans to react quickly. But how can we process visual input and decide on what to look within a matter of milliseconds? One of first models for visual attention introduced by James (1890) proposed that visual processing works by passing the input image first through a fast bottom-up step which is followed by a slow top-down one. The bottom-up pass detects regions, known as salient, which "pop out" from the receptive field (i.e., part of the neuron reacting to the stimulus). The top-down pass evaluates the content of the image (e.g., identify what kind of object is present and where it is located). Such a configuration allows for fast processing of the visual input as slower tasks are applied only on a fraction of the input image.

Since pioneering work by James (1890), the study of visual attention attracted the interest from multiple disciplines including like psychology, neuroscience, artificial intelligence, and computer vision. Because the problem is fundamentally connected with how the brain works a great deal of work was done by neuroscience community. On the other hand, artificial intelligence scientists focused on developing computational methods capable of recreating attention-driven maps taking place in the human brain. As for computer vision, saliency detection was found to be useful for a number of applications including: speeding up class-specific classifiers, content-aware image resizing, and collection summarization. In this survey, we focus on the evolution of saliency detection from a computer-vision perspective.

We divide computer-vision methods for visual saliency into three categories: psychologically-inspired, segmentation-based and data-driven. Psychologically-inspired methods are based on psychologically-plausible architectures of image processing in the brain. They usually consist of multiple feature maps that are combined under certain criteria into a final saliency map. Segmentation-based methods use cues such as texture and region similarity to separate redundant regions from salient ones. Data-driven methods use machine-learning techniques to create classifiers capable of separating salient locations in the image from the background.

Other reviews on visual attention exist under the point of view of neuroscience (Baluch and Itti, 2011; Carrasco, 2011), computer vision (Borji and Itti, 2013) and cognitive psychology (Bundesen and Habekost, 2005). For a computational-methods benchmark see Borji et al. (2012).
In this survey, we focus on saliency detection methods that try to identify regions within the image where the human eye is likely to focus first. This is a different perspective from saliency regions for image description as used in the work by Kadir and Brady (2001) and Kadir et al. (2004). Kadir et al. seek to find salient local regions in the image so that feature vectors calculated over such regions can be used for object recognition. We, on the other hand, concentrate on saliency detection for object localization.

2 Physiologically-inspired saliency detection

Early attention models for visual attention were built on results known from neuroscience and cognitive psychology.

2.1 Feature-integration theory

In cognitive psychology, Treisman and Gelade (1980) proposed the feature-integration theory of visual attention. They suggested that attention is based on a map of salient objects (i.e., a saliency map) which is built by integrating a set of calculated maps created from image features such as color, contrast, and orientation.

Building on the feature-integration theory, the active-vision system by J. and Ferrier (1988) calculated saliency map by combining (i.e., integrating) a number of filter-response maps for features such as color and line orientations. Here, the contribution of each feature to the overall saliency map was either amplified or inhibited depending on the context (e.g., for rectangle-like objects line feature would be amplified while color ones would be inhibited) by means of weights called amplification parameters.

Another key component of vision attention is visual acuity, which is the capacity to resolve two nearby objects (i.e., spatial resolution). Visual acuity is closely linked to eye movements. Eye movements help us refine the object’s location so that the object’s image becomes more clear as it projects onto the fovea (i.e., the area of the retina with the highest density of spatial and color sensors.)

Human-vision acuity is inversely proportional to the viewing angle, the property known as \( \frac{1}{\theta} \) law (see visualization in Fig. 1). As the result, locations with a large angle from the focus of attention are more blurry than ones with smaller angle. This relationship between attention
and viewing angle was used by Carlson et al. (1981) to develop a scale-invariant method for object recognition. The inverse-angle model was also used by Burt (1988), who created an approximate distribution of the $\frac{1}{\theta}$ measure using multi-resolution Gaussian pyramids. Here Burt (1988) showed that alternative distributions are not feasible biologically, and the $\frac{1}{\theta}$ distribution provides a balance between the two conflicting requirements of obtaining a wide field view and high acuity, both subject to constraint on the size of the optic nerve.

A direct consequence of the $\frac{1}{\theta}$ acuity model is that our visual perception cannot be simultaneously attentive at different spatially-disjoint locations, a property known as winner-take-all (Koch and Ullman, 1985), which we discuss next.

### 2.2 Winner-take-all

Given the focal nature of visual acuity, selective visual attention plays an important role in early visual processing in humans (Zeki, 1978; Zeki et al., 1991). Based on these results, Koch and Ullman (1985) proposed one of the earliest neurobiological model of visual attention. Their model consists of the following steps: (1) A number of feature maps are computed in the parallel, (2) Features maps are selectively combined to create a central saliency map,  

![Fig. 1: Visualization of the $\frac{1}{\theta}$ attention law. Notice how attention is concentrated on the sunflower in the top left corner, while parts with a higher viewing angle are more blurry.](image-url)
(3) The uniqueness of locations in the central map is used for attention selection, which is implemented using winner-take-all network. (4) Inhibition of visited locations cause an automatic shift towards the next most conspicuous location.

Almost in parallel to the neuroscience community, researchers in computer vision realized the need for selective visual-search mechanisms as a means for reducing the complexity of object-recognition tasks.

Work by Treisman and Gelade (1980); Tsotsos (1988, 1989) showed that visual-search tasks where target locations are given in advance can be computed in linear time in terms of the size of the image. In contrast, visual searches performed without explicit location information make the problem NP-complete.

The winner-take-all method by Itti et al. (1998); Itti and Koch (2000) builds on visual attention model introduced by Koch and Ullman (1985), and decomposes the input image into a set of feature maps that are subsequently combined into a master saliency map. Multi-scale feature maps are constructed from image statistics like contrast, color, and orientation. Contrast maps are constructed by applying low-pass filters on 9 different scales. The resulting maps preserve spatial discontinuities that stand out as salient. Final contrast maps are calculated as a difference between maps obtained from fine and coarse scales. This difference operations emulate "center-surround" behavior in visual receptive fields. Color information is incorporated in a similar manner, i.e., the pixel intensity and the three color channels (i.e., red, green, and blue) are used to calculate maps at different scales. Orientation information is incorporated by applying oriented filters (e.g., Gabor filters) at multiple scales. Again, orientation maps are calculated as differences of fine and coarse scales. Differences of contrast, color, and orientation maps create a number of feature maps. Maps are normalized and combined into the three categories. A final saliency map is a normalized average.

The integration of multi-scale feature maps to form global saliency maps and the winner-take-all principle allow for models of focal visitation of conspicuous locations. However, as the eyes move to fixate at different locations, what is the underlying perceived source of attention? Is visual attention drive by a combination of the perceived quality and quantity of information at certain locations in the image?
A Model of Saliency-Based Visual Attention

INTRODUCTION

Models of attention include "dynamic routing" models, in the basis of several models [5], [6]. It is related to the so-called plausible architecture, proposed by Koch and Ullman [4] and at scene-dependent) control [3], [2], [1].

Attention, which information from only a small region of the visual field can be selected through dynamic modifications of cortical connectivity or through the establishment of specific temporal patterns of activity, under both top-down (task-dependent) and bottom-up strategies [7]. Visual input is first decomposed into a set of topo-"feature integration theory," explaining human visual search in detail.

The model used here (Fig. 1) builds on a second biologically-
complex problem of scene understanding by rapidly selecting, in a top-down, volition-controlled, and task-dependent manner [2].

For information on obtaining reprints of this article, please send e-mail to:tpami@computer.org, and reference IEEECS Log Number 107349.

Fig. 2: Diagram for saliency detection method by Itti et al. (1998). Feature maps are calculated based on colors, intensity, and orientations on multiple scales which, after normalization are combined into a final saliency map. Salient locations are detected as local maxima in the saliency map.
2.3 Information maximization

Alternative line of work suggested that eyes first fixate on locations with the "most information" available in the scene. Lee and Yu (1999) looked at the fixation eye movements in the light of an information-theoretical framework. In their work, locations that are visited first are the ones with maximum uncertainty, modeled as entropy of the neuronal ensemble response of a hypercolumn in V1 (i.e., a region of the visual cortex). To explain "inhibition-of-return" principle entropy of the new location is discontinued by mutual information. Being able to explain both: pop-out and "inhibition-of-return" the framework proposed by Lee and Yu (1999) established new frontier for saliency research.

Work by Renninger et al. (2005) was the first to use the information-maximization framework of attention to implement a computer-vision system. For simplicity, authors worked with silhouettes (i.e., binary shapes), where silhouette was represented by set of points and tangent directions (i.e., edgelets). The method calculated a probability distribution of fixation locations, given histograms of edglets. Fixation points were chosen as locations with minimum entropy. The criterion for moving between fixation points was based on fact that there is little uncertainty over regions that were visited earlier (Fig. 3). Because the entropy of the current location is dependent on the history of prior fixation points, the same points are never attended twice.

Bruce and Tsotsos (2005) pointed out that self-information is more a more adequate metric for saliency detection than entropy. To illustrate the distinction, consider the image in Fig. 4. While the white region, which draws significant attention, has high self-information, it has low entropy. This can be explained by the fact that entropy is closer to the measure of local activity while self-information to local contrast. Bruce and Tsotsos (2005)' method works
Fig. 4: The white area being the most fixation location in the image has highest self-information but lowest entropy. Figure from Bruce and Tsotsos (2005).

by first applying independent component analysis (ICA) on a large sample of image patched to construct an independent basis. This basis is used to represent each image patch, and calculate self-information.

2.4 Spectral methods

Spectral methods, assume that a location is salient if it is a distinctive or non-redundant part of the image (Hou and Zhang, 2007). Barlow (1961) proposed that the visual system removes redundant visual information from processing. That is, features frequently occurring in the image are suppressed while the ones that deviate from the norm are treated more sensitively (Koch and Poggio, 1999). This idea inspired Hou and Zhang (2007) to find salient locations by removing redundancies from the image. They decomposed the image into innovation (i.e., the novel part in the image) and prior knowledge (i.e., a redundant part). Innovation is the part containing salient locations. The method proceeds by representing the image by a log spectrum, which is calculated using Fourier transformation. Hou and Zhang found that different images share similar log-spectra and that the plot of log intensity versus frequency is approximately linear. While frequent regions in the image are responsible for such shape, salient objects in the image are likely to cause discontinuities on it. Discontinuities are found from the spectral residual, which is calculated as a difference between the log spectrum and its average. The final saliency map is obtained by calculating the inverse Fourier transform on the residual.

Guo et al. (2008) argued that the spectral residual is not necessary for saliency computation. Instead, they proposed to use the phase spectrum of the Fourier Transform. The motivation for this is the fact that the amplitude spectrum is responsible for amplitude of each sinusoidal component, while the phase information specifies where each of of sinusoidal component
resides in the image. Consequently, phase information is useful for locating less homogeneous regions.

2.5 Context as a saliency cue

The idea to treat saliency in the global context was first introduced by Oliva et al. (2003) in their work on object detection. Their model assumed that the probability of a specific object being at a certain location depends on the object likelihood, saliency, and contextual priors. Oliva et al. showed that contextual priors calculated over the whole image can improve saliency detection, a claim largely supported by experiments (De Graef, 1992; Chun and Jiang, 1998; Henderson, 2003; Neider and Zelinsky, 2006). Experiments by Biederman et al. (1982) showed that if items are in physically impossible locations in the image their perception takes longer times for humans. Later work by Zhang et al. (2008) built on top of the model proposed by Oliva et al. (2003), but with probability distributions learnt on a separate training set of natural images. They showed that their method is more biologically plausible as organism learns to distinguish rare targets during their lifetime, while if only a single image is used such differentiation is impossible.

Goferman et al. (2010) extend the definition of saliency as an area in the image that represents a scene as opposed to just an object in the image. Contrary to the widespread definition where the method has to find the object independently of the scene, their method focuses on finding part of the image defining the scene. For example, Fig. 5 is the comparison between generic salient regions versus context saliency as defined by Goferman et al. (2010). It was shown that such an approach is preferable where the context of the object is as important as the object itself. Examples of applications include image re-targeting (i.e., re-sizing the image without losing its functional properties), image summarization (i.e., selection of few images from a collection to represent the collection).

3 Segmentation-based methods

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

Fig. 5: The difference between salient object with and without context. Top row: original images, second row: scene labels, third raw: generic saliency and last row context saliency (figure from Goferman et al. (2010))
Image segmentation and saliency detection solve similar problems. Whenever we look at an image of objects, regardless of those objects being natural or man made, we can observe various non-overlapping regions of homogeneous visual appearance (i.e., regions containing pixels of similar visual texture and colors). These regions or segments can be detected automatically by image-segmentation algorithms (Pal and Pal, 1993), which work by grouping or clustering those pixels having similar visual characteristics while also trying to create boundaries between dissimilar regions.

The grouping of image regions into objects of unique appearance underpins the segmentation-based methods for saliency detection. The methods we discuss in this section use image segmentation as a step to generating salient locations. Here, it is assumed that segmentation: (a) is object independent, (b) works on multiple scales, and (c) provides more accurate measurements given its free-form region boundaries in comparison to rectangular bounding boxes. These properties allow us to define "objectness" measures directly on image regions.

### 3.1 Global contrast saliency methods

Segmentation serves as essential step in most saliency detection methods as it allows to remove unnecessary detail as well as it allows to process the image on a regional level. One can define a salient object as the one that is the most different from the others. Cheng and Zhang (2011) employ such a strategy. After segmentation they define a saliency metric for a given region as a linear combination of the color distance to all other regions in the image with weights depending on spatial location. Color histograms are used as features for each region. Later, Perazzi et al. (2012) used the Euclidean distance as a metric for region comparison with spatial weights given by Gaussians. The key contribution in Perazzi et al. (2012) is the implementation, which, based on permutohedral lattice embedding introduced in Adams et al. (2010), allowed for the calculation of contrast in linear time in terms of the number of regions.

### 3.2 Saliency by composition

Another way we can define saliency is to consider it as a measure of how hard it is to compose the visual information within a bounding box by using visual information obtained from regions from outside it. This composition approach was used by Feng et al. (2011)’s method, which first divides the image into segments, and then calculates the cost of composing a bounding
box using outside regions as a linear combination of appearance and spatial proximity of regions. By using the efficient graph-based segmentation technique by Felzenszwalb and Huttenlocher (2004), Feng et al. (2011) reports a processing speed of 2 seconds per image.

### 3.3 Multi-cue contrast methods for object proposals

Because we usually do not know where objects of interest are located in images, it is common for object-detection methods to evaluate a classifier at rectangular regions centered at every pixel location in the image (i.e., sliding-window approach). Instead of scanning all pixels, a more efficient approach is to test only a small subset of bounding boxes that are deemed salient. This is the approach followed by selection-search methods.

Uijlings and van de Sande (2013) use graph-based segmentation (Felzenszwalb and Huttenlocher, 2004) to generate segments that are then iteratively merged to nearby similar regions until only one region remains (i.e., the object candidate or proposal). The criterion for merging regions uses a multi-component similarity function that is based on measurements of color, texture, size, and object’s fill. The size component of the cost-function encourages smaller regions to merge early while the object’s fill component discourages the merging of regions that are far apart from each other. In addition to using similarity of color and texture, the criterion used by Uijling and van de Sande for merging regions is also designed to diversify final object candidates. Here, selective search is applied on different complementary color spaces such as RGB, HSV, and LAB, and also applied on separate color channels within each color space, i.e., the RG channels of normalized RGB plus intensity, the hue channel from HSV.

Alexe et al. (2010, 2012) build their work on the principles that salient object has at least one of the next three properties: (i) well-defined closed boundary in space (ii) different appearance from its surroundings (iii) it is unique within the image and stands out as salient. Multiple saliency cues are used for detecting objects under each category. In order to detect unique regions in the image global saliency measure introduced by Hou and Zhang (2007) is used on multiple scales. The measure favors unique regions within the image. Given a bounding box, its multi-scale saliency (MS) is given as a sum of individual saliency values weighted by number of pixels higher than a certain threshold divided by total number of pixels in the box. The color-contrast (CC) cue measures the dissimilarity of a window and its enlarged surrounding area (Liu et al., 2007). It is computed by first resizing the box by a factor (which is learnt from the dataset). Similar to Liu et al. (2007), color contrast is given as $\chi^2$ kernel between
LAB color histograms of the box and its surroundings. Such a measure was inspired by the center-surround measures used in Itti and Koch (2001). Two measures are used to capture closed boundary within the bounding box: edge density (ED) and superpixel straddling (SS). Edge density is calculated by first shrinking the box by a factor (which is learnt from the data) and is given by the number of pixels of the boundary which considered edges divided by inner box perimeter. Superpixel Straddling cue is another way to capture closed boundary which uses superpixels segmentation (Felzenszwalb and Huttenlocher, 2004). The measure is designed so that bounding boxes that have fewer superpixels having part inside of the box and other part outside had higher values. A bounding box is said to straddle a superpixel if there is at least one pixel inside and one outside. The SS measures the degree by which all superpixels straddle the bounding box. Learning is performed in a Bayesian framework. Parameters are found by maximizing the posterior probability where priors are set to the relative frequency.

Rahtu et al. (2011) use cascades to speed up the generation of object proposals. The method starts by generating initial bounding boxes from superpixel segmentation and a prior distribution of bounding boxes in the image. A feature vector for each bounding box is computed using a modified pixel-straddling measure (Alexe et al., 2010), boundary edge distribution, and window symmetry. A modified max-margin structured SVM is used to ensure that only boxes with high ranks are advanced to higher layers of the cascade.

Although methods in Alexe et al. (2010); Rahtu et al. (2011) have high recall and precision they suffer from high time complexity. Manen et al. (2013) addresses this issue by using a randomized Prim’s algorithm. Usage of the randomized Prim’s algorithm serves as a way to diversify proposals and decrease computational complexity. In their work, the problem of sampling connected groups of superpixels (i.e., superpixels that make up an object) is cast as the sampling of partial spanning trees with high sum of edge weights. A superpixel-similarity measure is given as a probability distribution learnt on the features, based on color similarity, common-border ratio, and size. To generate an object proposal, a starting superpixel is chosen randomly and a partial spanning tree is “grown” based on edge similarity. SS

3.4 Graph-based objects proposals with ranking

Graph-based methods for visual attention segment the image into disjoint regions. Here, pixels are nodes of a graph and the nodes are connected by similarity. The method in Carreira and Sminchisescu (2012) follows segment-then-rank pipeline. At first, segmentation is cast as a max-flow problem, where the objective is to minimize an energy function consisting of unary
and pairwise potentials (Hochbaum, 2008). By varying the initial seeds and parameters of the unary potentials, multiple segmentations are obtained. A post-filtering step is performed to merge similar regions and delete trivial ones. At first, small segments are deleted, then remaining segments are ranked according to ratio cut (Wang and Siskind, 2003). Top 2,000 segments are retained and clustered while the rest is discarded. A final ranking of segments is performed to shuffle remaining segments so that more "object-like" ones are on top.

3.5 Hierarchical segmentation proposals with ranking

Hierarchical segmentation allows us to control the number of regions present and, as a result, can be used to diversify object proposals. Hoiem et al. (2011); Endres et al. (2014) calculates 4 successively coarser segmentations that are used to calculate probability distribution of occlusion figure/ground. After averaging across segmentations, the distribution is used to obtain agglomerative grouping using method from Arbeláez et al. (2011). Object proposals are generated starting from seeds that are chosen from hierarchical segmentation so that regions around them are large enough to calculate reliable color and texture distributions. Proposals are inferred from foreground/background labeling which is computed by minimizing an energy function proportional to the unary and pairwise potentials. The unary potential, called superpixel-affinity term, is an image statistic calculated from a superpixel region. It represents how likely is that region to lie on the same object as the seed. Region features like color, texture, boundary strength, and layout agreement are used to learn foreground probability distribution using boosted decision-tree classifier. Pairwise potentials enforce cost for assigning different labels for adjacent superpixels if the boundary between them is weak. In order to generate a large set of object proposals, conditional random fields (CRF) learning is performed with multiple sets of parameters. After proposals are generated, they are ranked using a ranked structured support vector machine.

4 Data-driven methods

Recently, saliency detection methods that use supervised machine-learning techniques have emerged. As saliency ground-truth data is hard to obtain, the problem of saliency detection is replaced with a similar one: proposal generation. Proposal generation solves the problem of localizing a subset of bounding boxes in the image which are likely to contain a salient object. Such a problem lies in between object localization and saliency detection. Starting
from early supervised methods that used SVMs (Kienzle et al., 2006), data-driven methods gained momentum and hold most of the state-of-the-art results on most datasets.

4.1 SVMs for saliency detection

A main critique of psychologically-inspired methods is their reliance on manually engineered feature maps which in turn require the settings of many parameters such as the size and shape of filters. In contrast, data-driven methods attempt to learn most of these parameters from the data. Kienzle et al. (2006) was the first to apply SVMs to detect salient regions in the image. A binary SVM classifier was trained on pixel features to differentiate salient locations from non-salient ones. Kienzle et al. (2006) trains the classifier with image regions extracted at locations provided by eye-tracking fixation locations. Here, negative examples in the dataset were chosen as a set of non-fixation locations or background. Although the saliency measure is given as a non-linear combination of linear filters, this approach is non-parametric as a choice of filters (i.e., images themselves) is data-dependent.

The notion of visual saliency of objects in images has been used for obtaining faster object detection. This type of saliency is also called "objectness". Cheng et al. (2014) tried to speed object detection by finding only a subset of bounding boxes for classifier evaluation. It was built on the principles that good generic objectness score has to satisfy the following criteria: (i) Achieve high detection rate (i.e., miss as little as possible potential objects ) (ii) Generate small number of object proposals (iii) Be fast to evaluate (iv) Generalize well to unseen objects.

Cheng et al. (2014) noticed that the bounding boxes of objects, when resized to small size (e.g., $8 \times 8$ pixels) are similar to each other regardless of their initial size. Thus, they proposed to use normed gradients of the resized image as a feature vector. The method is a two-step procedure where the first step is to learn a linear SVM using normed gradients, and the second is to learn the objectness measure using the bounding box size. The SVM is trained using ground truth bounding boxes as positive examples and randomly sampled boxes as negative. Each image is resized to 36 predefined sizes so that proposal generation was possible on multiple scales. Finally, in order to calculate the objectness score, the size of the bounding box is used as the SVM parameters for ranking proposals. Cheng et al. (2014) report that on average it takes about 0.003 second per image to generate and rank object proposals which is the fastest method by far. This method’s speed is around 1000x faster than the methods by Alexe et al. (2012) and Zhang et al. (2011) without sacrificing accuracy.
4.2 Low-rank factorization

Factorizing an image into salient and redundant parts can be achieved using low-rank factorization algorithms (Candès and Recht, 2009). Low-rank factorization decomposes the image matrix into a low-rank (i.e., redundant) and a sparse (i.e., salient) matrices. Unfortunately, if applied directly on the image pixels, the sparse component will consist only of the image noise as opposed to actual salient regions. Yan et al. (2010) addressed this problem by first learning an overcomplete sparse-code dictionary (Mairal et al., 2010), a strategy used in the brain Olshausen and Field (1996), and then applying low-rank factorization to the new feature space. Later, Shen and Wu (2012) pointed out that due to small size of patches (i.e., patches of size $8 \times 8$ were used in Yan et al. (2010)) which make up a salient object salient patches will not be sparse anymore. Instead, they proposed to learn a linear transformation of the feature space first, and only then, apply low-rank decomposition. Furthermore, to avoid overfitting to training examples Zou et al. (2013) proposed an unsupervised method for saliency detection with segmentation priors. The idea behind segmentation priors is that background regions have high probability of connecting to the border. Thus, coarse segmentations are used to weight down segments whose perimeters intersect with the background. Once such priors are learnt, they are used to transform feature matrix to be used for low-rank decomposition. To further refine the saliency map, post-smoothing step is applied to penalize high-saliency differences between nearby pixels.

4.3 Deep learning

Recently deep learning was established as a state-of-the-art paradigm for object-recognition achieving top results on most large-scale datasets Krizhevsky et al. (2012), thus it was a matter of time before it was applied to saliency detection. In Erhan et al. (2014) the problem of saliency detection is cast as regression using deep neural networks with modified loss. The loss learns predictors for locations of the corners of the bounding box during training. Due to high representation capability of the deep neural networks authors report that few bounding boxes per image (10 compared to 1000 in previous methods) were enough to achieve state-of-the-art results on PASCAL VOC2007 dataset.

Work by Wang et al. (2014) applied deep convnets for the task of joint saliency selection and segmentation. The method consists of two multi-layer convolutional networks: one for localization and another for segmentation, where both networks collaborate via latent
variables. Novel optimization of the model is achieved using EM-like algorithm where on the first step latent variables are updated using MCMC-based sampling, and on the second using back propagation with latent variables fixed.

4.4 Other approaches

Saliency detection in Liu et al. (2007) is cast as a binary labeling problem, which separates the salient object from the background. Liu et al. (2007) pointed out that saliency methods that only depend on local features assign high-saliency values to locations with high contrast. Although contrast is a good cue for saliency, high contrast does not guarantee salient object. To address this issue, Liu et al. (2007) extended the set of features to include region and global image statistics. The problem is formulated in the Conditional Random Field (CRF) framework, where CRF models the probability of each pixel being salient or not. The energy function, which is proportional to a conditional distribution, is a sum of unary and pairwise terms. The unary term is a linear combination of independent image features. The pairwise term models spatial relationship between nearby pixels and is set to discourage pixels having different labels. Weights of the linear combination of image features are learnt via maximum-likelihood principle.

Zheng and Hua (2011) extented Liu et al. (2007) by adding extra "auto-context" term into energy function (Tu, 2008). Auto-context allows to learn posterior probability distributions on image patches with non-local information. Learning is performed with multi-layer boosting classifier which allows to fuse local image information with context. Zheng and Hua (2011) introduced an optimization technique that jointly optimizes segmentation, saliency, and auto-context.

Zitnick and Doll (2014) built on the observation that the more the number of contours wholly contained in the bounding box the more likely that the bounding box contains an object. Computationally Zitnick and Doll (2014) implemented this observation using edges. In their approach objectness score measures the number of edges in the bounding box minus edges of the contours straddling it. For that purpose edges are computed using computationally efficient algorithm Dollar and Zitnick (2013). Then, individual edges are clustered into groups for which affinity is defined in terms of relative position and orientation. Objectness score is defined as a sum of edge strength of all edge groups within the box minus the ones straddling it. It was also reported that final implementation is fast: the method takes under a second to find salient locations.
In Jiang and Davis (2013) the problem of saliency detection is formulated as the one of \textit{facility location}. Originated from operations research, facility location is the problem to set up a number of facilities which would maximizing the profit from a set of clients. To find saliency initial pool of clients is calculated from superpixel segmentation. Since each location has an associated cost to the closest facility the total profit is a modular function defined on a harmonic graph (Sviridenko, 2002). One of contributions in Jiang and Davis (2013) is the greedy optimization procedure which guarantees approximately optimal solution. Then, final facilities (salient locations) are given by a solution to facility location problem which are further refined with high and low-level priors.

Under assumption that attention is initially attracted to specific locations in the image and only then propagated to other regions Liu et al. (2014) proposed to model such process with Linear Elliptic partial differential equation (PDE) with Dirichlet boundary conditions. Solution to the PDE, known as guidance map, together with other priors are used to select saliency seeds, which are sufficient to reconstruct saliency map.

5 Challdges and future work

5.1 Challenges and open issues

Although the problem of saliency detection was revisited multiple times in recent years in the form of object proposals challenges and open issues still exist.

Originally object proposals were created as a way to speed-up expensive-to-evaluate supervised classifiers, but how does it differentiate them from cascade methods? In classical cascade methods (Dollár et al., 2012) cascade classifiers are used to prune bounding boxes earlier in the cascade, while the methods for objects proposals are designed to be independent of the classifier. So are object proposals necessary just to make object detection faster? What will be their use once computational power increases? The role of object proposals will depend on their utility. Currently methods claim that they are able to generalize to unseen objects, but it still remains unclear how can this be evaluated.

Evaluation criteria is another challenge in the study of object proposals. How should classes be divided to establish fair class-independent evaluation? How many proposals to generate per image and how much of an overlap between bounding boxes to treat detection as successful? Since different methods are tailored for specific applications setting the baseline is important
for the progress.

Study of robustness for object proposals is important in showing their utility. Currently methods do not include experiments on robustness when images are perturbed with noise. Repeatability is another property to be expected from object proposals: a good saliency detection method is expected to generate similar sets of bounding boxes from slightly transformed images.

## 5.2 Future work

There are many applications which could benefit from object proposal generation. For example, it was shown in (Alexe et al., 2012) that object proposals can speed-up object localization process, but can it also increase the accuracy? In object localization negative examples are usually sampled from same images containing positive ones but with restriction that there is small overlap with ground truth. Object proposals, being able to generate object-like bounding boxes, could be used to generate a better negative examples. Furthermore, the ability to separate object from the background could be useful for unsupervised or semi-supervised applications.

In model-free tracking, where the object being tracked is not known in advance, objectness prior could be used to improve tracking accuracy. One of the main challenges in tracking is the drift - the process when the tracker accumulates the error until the object is lost completely. If, in addition to the tracker, there is a knowledge of how likely the bounding box to contain any object the drift effect can be decreased. As for unsupervised methods, object proposals could be used on a preprocessing step.

In weekly-supervised action or interaction recognition in videos preprocessing step usually involves localization of the action in time and space. This step can be made more accurate with the help of object proposals. Because the data usually involves dealing with videos the methods for extracting proposals should be generalized to videos.

Extending object proposals to the temporal domain is another direction worth consideration. Current methods could be trivially extended to video setting if applied on different frames independently, but between frames correlation could be used to improve the quality and the speed of the proposals.
6 Conclusions

In this paper we discussed the evolution of computational methods on saliency detection. This was done by grouping methods into three different categories: psychologically-inspired, segmentation-based and data-driven. For each category we provided overview of the methods, explaining how each moved the field further. Similar problem of object proposals was discussed in the context of object localization. Being in the middle of saliency detection, segmentation and object localization object proposals show great promise as many applications are possible. Such application span supervised to unsupervised problems like object localization and tracking. Nevertheless scientific community needs to focus on problems like robustness, repeatability and purpose of the object proposals to see the extend of which object proposals should be used.

References


