A saliency map based on sampling an image into random rectangular regions of interest

Tadmeri Narayan Vikram\textsuperscript{a,b,}\textsuperscript{*}, Marko Tscherepanow\textsuperscript{a}, Britta Wrede\textsuperscript{a,b}

\textsuperscript{a}Applied Informatics Group, Bielefeld University, Bielefeld 33615, Germany
\textsuperscript{b}Research Institute for Cognition and Robotics (CoR Lab), Bielefeld University, Bielefeld 33615, Germany

\textbf{A R T I C L E  I N F O}

Available online 22 February 2012

Keywords:
Visual attention
Saliency map
Salient region detection
Eye-gaze prediction
Bottom-up attention

\textbf{A B S T R A C T}

In this article we propose a novel approach to compute an image saliency map based on computing local saliencies over random rectangular regions of interest. Unlike many of the existing methods, the proposed approach does not require any training bases, operates on the image at the original scale and has only a single parameter which requires tuning. It has been tested on the two distinct tasks of salient region detection (using MSR A dataset) and eye gaze prediction (using York University and MIT datasets). The proposed method achieves state-of-the-art performance on the eye gaze prediction task as compared with nine other state-of-the-art methods.

\textcopyright 2012 Elsevier Ltd. All rights reserved.

1. Introduction

Visual attention and its associated cognitive neurobiology has been a subject of intense research over the last five decades. The attention mechanism in humans helps them in focusing their limited cognitive resources to the context relevant stimuli while suppressing the ones which are not important. Having an artificial system which could simulate this attention mechanism could positively impact upon the way in which various cognitive and interactive systems are designed. In view of this, the computer science community has invested considerable time and efforts to realize a computational model of visual attention which at least partially exhibits the characteristics of a human visual attention system. Several computational approaches to visual attention have been proposed in the literature since the seminal work of Koch and Ullman in 1985 [1]. The same authors proposed the concept of a saliency map, which constitutes the surrogate representation of the visual attention span over an input image for a free viewing task. Saliency maps are now employed for several applications in robotics, computer vision and data transmission. Some of the successful applications are object detection [2,3], object recognition [4], image summarization [5], image segmentation [6,7], image compression [8] and attention guidance for human–robot interaction [9,10]. Most of the existing visual attention approaches process pixels or image windows sequentially, which is in contrast to the way the human visual system operates. It has been shown in the literature that the receptive fields inside the retina operate randomly at both positional and scale spaces while processing the visual stimuli [11].

Motivated by this, we propose a simple and efficient method which is based on random sampling of the image into rectangular regions of interest and computing local saliency values. The proposed method is shown to have comparable performance with existing methods for the task of salient region detection. Furthermore, the proposed method also achieves state-of-the-performance for the task of predicting eye-gaze. The method has been tested on MSR A [12], York University [13], MIT [14] datasets and benchmarked along with nine of the other existing state-of-the-art methods to corroborate its performance.

The paper is organized as follows. We provide a review of the existing computational approaches to visual attention with a brief description of their strengths and shortcomings in Section 2. The motivation leading to the current work is described in Section 3. The proposed approach is described by means of a pseudo-code and a graphical illustration in Section 4. In Section 5, we describe the experiments carried out to validate the performance of the proposed method for the tasks of eye-gaze prediction and salient region detection. Finally, we conclude this work by highlighting its current shortcomings with a brief discussion about the future directions of the current work in Section 6.

2. Literature review

The theoretical framework for the computation of saliency maps was first proposed in [1] and later realized in [15].
Subsequently it has led to the development of several other saliency approaches based on different mathematical and computational paradigms. The existing approaches can be classified into seven distinct categories based on the computational scheme they employ.

- **Hierarchical approaches**: They perform a multi-scale image processing and aggregate the inputs across different scales to compute the final saliency map.
- **Spectral approaches**: They operate by decomposing the input image into Fourier or Gabor spectrum channels and obtain the saliency maps by selecting the prominent spectral coefficients.
- **Power law based approaches**: These approaches compute saliency maps by removing redundant patterns based on their frequency of occurrence. Rarely occurring patterns are considered salient while frequently occurring patterns are labeled redundant.
- **Image contrast approaches**: The mean pixel intensity value of the entire image or of a specified sub-window is utilized to compute the contrast of each pixel in the image. The contrast is analogously treated as the image saliency.
- **Entropy-based approaches**: The mutual information between patterns is employed to optimize the entropy value, where a larger entropy value indicates that a given pattern is salient.
- **Center-surround approaches**: These approaches compute the saliency of a pixel by contrasting the image features within a window centered on it.
- **Hybrid approaches**: Models of this paradigm employ a classifier in combination with one or more approaches to compute saliency.

We briefly explore the existing saliency approaches based on the aforementioned categories in the sub-sections to follow. Interested readers are pointed to [16,17] for a more detailed and exhaustive review on saliency approaches.

### 2.1. Hierarchical approaches

The most popular approach in this category is the one proposed by Itti et al. [15]. This approach computes 41 different feature maps for a given input image based on color, texture, gradient and orientation information. The resulting feature maps are fused into a single map using a winner takes all (WTA) network and an inhibition of return mechanism (IOR). This approach has served as a classical benchmark system. Despite its theoretic appeal, the Weibull's distribution based saliency map has been proposed in [24], and its performance on the eye-gaze data is shown to be better than other power law based approaches. The method employs maximum likelihood estimates (MLE) to set the parameters of various distributions. Despite its theoretic appeal, the Weibull's distribution based saliency map has a large parameter set and the heuristic to fix and optimize them constitutes a major drawback.

### 2.2. Spectral approaches

Several approaches to compute a saliency map which operate at the spectral domain have thus been proposed in the literature. A Gabor wavelet analysis and multi-scale image processing based approach was proposed in [18]. Fourier transform has also been extensively utilized to generate saliency maps as seen from [8,19]. In the approach of [8], the input image is decomposed into four different channels based on opponent colors and intensity. A quaternion Fourier analysis of the decomposed channels is carried out to compute the master saliency map. Two dimensional Gabor wavelets of Fourier spectrum have also been utilized for computing saliency maps in [20] and the final saliency map is re-weighted using a center-bias matrix. Spectral residual approaches proposed in [21,22] are another class of approaches which highlight the saliency by removing redundancy from the Fourier spectrum. The saliency map is thus constructed from a residual frequency spectrum obtained by the difference between an image frequency spectrum and a mean frequency spectrum. Though straightforward, the method is compounded due to the fact that the pattern of mean frequency spectrum is subjective. These methods are found to be successful in detecting proto-objects in images. As it can be seen from the illustrations in [8], Fourier-based methods are affected by the number of coefficients selected for image reconstruction and the scale at which the input image is processed. Like subspace analysis, the method results in the loss of information during image reconstruction and is compromised by illumination, noise and other image artifacts.

### 2.3. Power law based approaches

Power laws are basically used to model human visual perception. Several approaches to saliency which employ power laws have thus been proposed. A Zipf's law based saliency map has been proposed in [23]. This work is similar to those of [21,22], except that it operates directly on the pixels rather than on the spectral space. An integrated Weibull's distribution based saliency map has been proposed in [24], and its performance on the eye-gaze data is shown to be better than other power law based approaches. The method employs maximum likelihood estimates (MLE) to set the parameters of various distributions. Despite its theoretic appeal, the Weibull's distribution based saliency map has a large parameter set and the heuristic to fix and optimize them constitutes a major drawback.

### 2.4. Image contrast approaches

A simple but efficient approach based on computing the absolute difference of a pixel and the image mean was proposed in [25]. Similar approaches proposed by the same authors in [26,7] employ maximal symmetric regions and novel dissimilarity metrics respectively to compute the final saliency map. Another symmetry based saliency approach was proposed in [27], which employs color, orientation and contrast symmetry features to generate the final saliency map. It also reasonably simulates eye-gaze for free viewing task in case of natural sceneries with no specific salient objects in it. A distance transform based approach was proposed in [28], which computes an edge map for each grayscale threshold and fuses them to generate a final saliency map. These methods are successfully applied for proto-object detection and out-beats many state-of-the-art methods without having many of their drawbacks. Poor global contrast of an image affects the performance of these methods and local-statistics based approaches for saliency computation have been proposed in the literature to address this problem.

### 2.5. Entropy-based approaches

The popular approach of [29] is based on local contrasts and maximizes the mutual information between features by employing independent component analysis (ICA) bases. A set of ICA bases is pre-computed using a patch size of 7 × 7 pixels. Subsequently it is used to compute the conditional and joint distribution of features for information maximization. Experiments conducted on the York University eye-gaze dataset [13] have proven its efficiency. But this method is constrained by its emphasis on edges and neglects salient regions [30]. It also adds a spurious border effect to the resultant
image and requires re-scaling of the original image to a lower scale in order to make the computational process more tractable. Another ICA based approach was proposed in [31] where image self-information is utilized to estimate the probability of a target at each pixel position. It is further fused with top-down features derived from ICA bases to build the final saliency map. The method proposed in [32] employs sparse bases to extract sub-band features from an image. The mutual information between the sub-band features is calculated by realizing a random-walk on them and initializing the site entropy rate as the weight of the edges in the graph. An extension of this paradigm can be seen in the recent approach proposed in [33], where entropy of a center versus a surround region is computed as the saliency value of a pixel. Other entropy-based approaches like [34,35] employ incremental coding length to compute the final saliency map. These methods which rely on information theoretic approaches are in general constrained by the requirements of training bases, the patch size parameters and the size of the training bases.

2.6. Center-surround approaches

A saliency map based on discriminant center-surround entropy contrast which does not require training bases was proposed in [36]. It correlates well with human eye-gaze data but is constrained by the subjectivity involved in the computation of weights for fusing the different maps to compute the master saliency map. A recent center-surround contrast method [37], identifies pre-attentive segments and computes the mutual saliency between them. An innovative application of this saliency map has been found useful for pedestrian detection. A multi-scale center-surround saliency map was proposed in the work of [38], where the saliency of a pixel is determined by the dissimilarity of the center to its surround over multiple scales. Local analysis of gradients along with center-surround paradigm is advocated as an alternative to multi-scale image analysis as it can be seen in [39,40]. In [39], local steering kernels are employed to compute the saliency of a pixel by contrasting the gradient covariance of a surrounding region. The method is shown to be robust to noise and drastic illumination changes, but it is computationally expensive as a set of compound features needs to be computed for each pixel in the image. In order to achieve a tractable run-time, the image is down-scaled. The approach proposed in [40] uses color and edge orientation information to compute the saliency map, using a framework similar to the one presented in [39]. Selecting an appropriate window size for center and surround patches plays an important role in obtaining a higher quality saliency map.

2.7. Hybrid approaches

To overcome the issue of patch size parameters, many machine learning based saliency approaches have been proposed. A graph-based visual saliency approach is proposed in [41] which implements a Markovian representation of feature maps and utilizes a psychovisual contrast measure to compute the dissimilarities between features. A saliency map based on modeling eye-movements using support vector machines (SVMs) is proposed in [42]. However, this approach requires enormous amounts of data to learn saliency weights reasonably. Another successful approach based on kernel density estimation (KDE) is proposed in [43]. The authors recommend to construct a set of KDE approaches based on region segmentation using the mean shift algorithm. Based on its color distinctiveness and spatial distribution, the color saliency and spatial saliency of each KDE approach are evaluated. The method is found to be successful in salient object detection, but has unwieldy run-time as it relies on image segmentation. A salient object selection mechanism based on an oscillatory correlation approach was proposed in [44]. The method outputs an object saliency map, which resembles image segmentation. The method requires fine tuning of neural network weights for an inhibitory mechanism which is heuristic and hence cannot be used for real-time visual attention. A Hebbain neural network based saliency mechanism which simulates lateral surround inhibition is proposed in [45]. Unlike [44], the method in [45] is computationally efficient as it employs the pulsed cosine transform to produce the final saliency map.

A color saliency boosting function is introduced in [46] which is obtained from an isosalient image surface. The final saliency map is based on the statistics of color image derivatives and employs an information theoretic approach to boost the color information content of an image. An extension of this work can be found in [47]. It performs reasonably well on both eye-gaze fixation and salient object detection datasets. In [48], the probability distribution of colors and orientations of an image is computed and either of these features are selected to compute the final saliency map. This method avoids multi-scale image analysis but fails when the image has a low color or gradient contrast.

A few hybrid approaches attempt to combine the positive aspects of multi-scale analysis, sub-band decomposition, color boosting and as well as center-surround paradigms. Fusion of the center-surround hypothesis and oriented sub-band decompositon to develop a saliency map has been proposed in [49]. A radically different approach which tries to detect salient regions by estimating the probability of detecting an object in a given sliding window is proposed in [50]. They employ the concept of super pixel straddling, coupled with edge density histograms, color contrasts and the saliency map of [15]. A linear classifier is trained on an image dataset to build a bag-of-features to arrive at a prior for an object in an image. The method is theoretically very attractive, but is subjected to high variations in the performance as too many features, maps and parameters are involved which require fine tuning.

3. Motivation

As it can be observed from our previous review, many of the existing computational approaches to visual attention are constrained by one or more of the following issues like : requirement of training bases, large set of tunable parameters, integration of complex classifiers, downsampling of the original image in order to obtain a tractable computational run-time, high, complexity associated with programming, and not being biologically motivated. Most of the existing approaches to compute saliency maps process input image pixels sequentially with fixed sliding windows. But salient regions or objects can occur at arbitrary positions, shapes and scales in an image. Research in vision sciences has shown that the human visual system has its receptive fields scattered randomly and does not process input stimuli in a sequential manner [11,51]. Furthermore, it has been shown in [52] that each visual stimulus is biased by every other stimulus present in the attention space.

In view of the aforementioned discussion, we propose a random center-surround method which operates by computing local saliencies over random regions of an image. This captures local contrasts unlike the global methods for computing saliency maps. Furthermore, it does not require any training priors and has only a single parameter which needs tuning. The proposed method also avoids competition between multiple saliency maps which is in contrast to the approaches proposed by [15,41].
4. The proposed saliency map

We consider a scenario where the input $I$ is a color image of dimension $r \times c \times 3$, where $r$ and $c$ are the number of rows and columns respectively. The input image $I$ is initially subjected to a Gaussian filter in order to remove noise and abrupt onsets. It is further converted into the CIE1976 $L^a*b^b$ space and decomposed into the respective $L^a$, $a^b$ and $b^b$ channels each of dimension $r \times c$.

The $L^a*b^b$ space is preferred over the other color spaces because of its similarity to the human psycho-sensory color space [26] and is also recommended as a standard to compute saliency maps in [30,25,26,7]. The saliency maps due to $L^a$, $a^b$ and $b^b$ channels are referred to as $S^L$, $S^a$ and $S^b$ respectively and their dimensions are also equal to $r \times c$.

Furthermore, $n$ random sub-windows are generated over each of the $L^a$, $a^b$ and $b^b$ channels. The saliency value at a particular coordinate position in a given channel is defined as the sum of the absolute differences of the pixel intensity value to the mean intensity values of the random sub-windows in which it is contained. The final saliency map $S$ is also of dimension $r \times c$. Computing the final saliency value for a given position as the Euclidean norm of saliencies over different channels in the $L^a*b^b$ space is recommended as an ideal fusion rule in [25,26,30].

A discrete uniform probability distribution function is used to generate the random sub-window coordinates, as this helps in placing windows without having any bias towards a specific window size or spatial region of the image. This property is important because salient regions or objects in an image can occur at arbitrary positions and scales.

The resulting saliency map $S$ is normalized in the interval $[0,255]$ and subsequently subjected to median filtering. The median filter is chosen because of its ability to preserve edges despite eliminating noise. To further enhance the contrast of $S$, we subject it to histogram equalization as recommended in [36,53]. This image enhancement is in consonance with the fact that the human visual system enhances the perceptual contrast of the salient stimulus in a visual scene [54,55].

Note that our approach does not downscale the input image to a lower resolution like other approaches [41,15,39]. In addition, our method does not require prior training bases in contrast to [31,13]. Also $n$ is the only parameter which requires tuning, the details of which are presented in the next section. The algorithm for the proposed approach is as follows.

Algorithm: Random center-surround saliency.

Input: (1) $I$ (RGB Image) of size $r \times c \times 3$
(2) $n$—Number of random windows

Output: $S$—Saliency map of size $r \times c$

Method

Step 1 : Apply Gaussian filter on $I$

Step 2 : Convert input $I$ to $L^a*b^b$ space

Step 3 : Generate random window co-ordinates $(x_i,y_i,x_j,y_j)$ = Generate_Windows ($n,r,c$)

Step 4 : Generate saliencies for each component
$S^L=\text{CompSal}(L^a,n,x_i,y_j,x_j,y_j)$
$S^a=\text{CompSal}(a^b,n,x_i,y_j,x_j,y_j)$
$S^b=\text{CompSal}(b^b,n,x_i,y_j,x_j,y_j)$

Step 5 : Compute $S$ by pixel-wise Euclidean norm
$S = \text{Fusion}(S^L, S^a, S^b)$

Step 6 : Apply median filter on $S$

Step 7 : Normalize $S$ in $[0,255]$

Step 8 : Apply histogram equalization to $S$

Algorithm: CompSal

Input: (1) $I$ (Gray scale Image) of size $r \times c$
(2) $n$—Number of random windows
(3) $c$—Number of columns

Output: $S$—Interim saliency map of size $r \times c$

Method

Step 1 : Set all elements of $S$ to 0

Step 2 : Update $S$

for $i = 1$ to $n$

$A_{i} = (x_{i} - x_{i-1}) \cdot (y_{i} - y_{i-1})$
$S = S + A_{i}$

Step 3 : End

Algorithm: Generate_Windows

Input: (1) $n$—Number of random windows
(2) $r$—Number of rows
(3) $c$—Number of columns

Output: $x_i,y_i,x_j,y_j$ each of size $n$

Method

Step 1 : Generate random window co-ordinates

Set all elements of $x_i,y_i,x_j,y_j$ to 0

$x_{i_1}$ = Random number in $[1,r-1]$
$y_{i_1}$ = Random number in $[1,c-1]$
$x_{i_2}$ = Random number in $[x_{i_1}+1,c]$
$y_{i_2}$ = Random number in $[y_{i_1}+1,c]$

Step 6 : End

Algorithm: Fusion

Input: (1) Matrices $A$, $B$ and $C$ of size $r \times c$

Output: $S$—Fused matrix $FM$ of size $r \times c$

Method

Step 1 : Apply pixel-wise Euclidean norm

Set all elements of $FM$ to 0

$FM_{ij} = \sqrt{A_{ij} \cdot A_{ij} + B_{ij} \cdot B_{ij} + C_{ij} \cdot C_{ij}}$

Step 7 : End

An illustration of the above paradigm is given in Fig. 1.

5. Experimental results

Several experiments were conducted to validate the performance of the proposed saliency map for two distinct tasks of...
eye-gaze prediction and salient region detection in free viewing conditions. Salient region detection and eye-gaze prediction are the two most significant applications of saliency maps. Salient region detection is relevant in the context of computer vision tasks like object detection, object localization and object tracking in videos [25]. Automatic prediction of eye-gaze is important in the context of image aesthetics, image quality assessment, human–robot interaction and other tasks which involve detecting image regions that are semantically interesting [14]. The contemporary saliency maps are either employed to detect salient regions as in the case of [30,25,28], or are used to predict eye-gaze patterns as in [31,39]. Only few of the existing saliency approaches like [41,15] have consistent performance on both of these tasks. Although these two tasks appear similar, there are subtle differences between them. Salient regions of an image are those which are visually interesting. But human eye-gaze which focuses mainly on salient regions are also distracted by semantically relevant regions [56]. The performance on the eye-gaze prediction task was validated on two different datasets of York University [13] and MIT [14]. Subsequent experiments to corroborate the performance on salient object detection task were conducted on the popular MSRA dataset [12].

The following parameter settings were used as a standard for all the experiments carried out. A rotational symmetric Gaussian low pass filter (size $3 \times 3$ with $\sigma = 0.5$; the default Matlab configuration) was used as a pre-processor on the images for noise removal as recommended in [25,26]. The number of distinct random sub-windows $n$ was set to $0.02 \times t \times c$, as it led to a more stable performance. Details about fine tuning $n$ is given in Section 5.1. A median filter of size $11 \times 11$ was employed to smooth the resultant saliency map before being enhanced by histogram equalization. The employed Gaussian filter, median filter and histogram equalization are based on the usual straight forward methods. All experiments were conducted using Matlab v7.10.0 (R2010a) on an Intel Core 2 Duo processor with Ubuntu 10.04.1 LTS (Lucid Lynx) as operating system. The inbulit srgb2lab Matlab routine was used to convert the input image from RGB colorspace to $L^*a^*b^*$ colorspace. This results in $L^*$, $a^*$ and $b^*$ images whose pixel intensity values are normalized in the range of $[0, 255]$. We selected nine state-of-the-art methods of computing saliency maps to compare and contrast the proposed method. The methods are global contrast2 [25], entropy3 [29], graphical approach4 [41], multi-scale5 [15], local steering kernel6 [39], distance transform7 [28], self-Information8 [31], weighted contrast9 [30] and symmetric contrast10 [26] based image saliency approaches.

The experiments conducted to corroborate the performance of the proposed method for eye-gaze correlation in a free viewing task is described in the following sub-section.

5.1. Experiments on eye-gaze prediction task

In order to empirically evaluate the performance on eye-gaze correlation task, we have employed the receiver operating characteristic (ROC)–area under the curve (AUC) as a benchmarking metric. Several popular and recent works like [41,42,49,57,20,30] employ the ROC–AUC metric to evaluate eye-gaze fixation correlation. An ROC graph is a general technique for visualizing, ranking and selecting classifiers based on their performance [58]. The ROC graphs are two-dimensional graphs in which the true positive rate (TPR) is plotted on the Y axis and the false positive rate (FPR) rate is plotted on the X axis. The TPR (also

---

**Fig. 1.** An illustration of the proposed method. The input image is subjected to Gaussian filter in the first stage. Subsequently it is converted into the $L^*a^*b^*$ space and the individual $L^*$, $a^*$ and $b^*$ channels are obtained. For the sake of simplicity we have considered three random regions of interest (ROI) on the respective $L^*$, $a^*$ and $b^*$ channels. Local saliencies are computed over each of these ROIs and the channel specific saliency maps ($S^*$, $S^a$ and $S^b$) are updated. The final saliency map is then computed by fusing the channel specific saliency maps by a pixel-wise Euclidean norm.
called hit rate and recall) and FPR (also called false alarm rate) metrics are computed as in [30,32]:

\[
\text{TPR} = \frac{tp}{tp + fn} \quad (1)
\]

\[
\text{FPR} = \frac{fp}{fp + tn} \quad (2)
\]

where \(tp\) is the number of true positives, \(tn\) is the number of true negatives, \(fp\) is the number of false positives and \(fn\) is the number of false negatives.

An ROC graph depicts relative trade-offs between benefits (TPR) and costs (FPR). Since the AUC is a portion of the area of the unit square, its value will always be between 0 and 1.0. An ideal classifier would give an AUC of 1.0 while random guessing produces an AUC of less than 0.5 [59]. The saliency map and the corresponding ground truth fixation density map are binarized at each discrete threshold in [0, 255]. This results in a predicted binary mask (from the saliency map) and a ground truth binary mask (from the fixation density map) for each binarizing threshold. The TPR and FPR for each threshold are subsequently computed. The ROC curve is generated by plotting the obtained FPRs versus TPRs and the AUC is calculated. In our case, the AUC indicates how well the saliency map predicts actual human eye fixations. This measurement also has the desired characteristic of transformation invariance, in that the ROC–AUC does not change when applying any monotonically increasing function (such as logarithm) to the saliency measure [31].

We have considered two popular datasets of York University [13] and MIT [14] for the eye-gaze prediction task. The York University [13] dataset contains eye fixation records from 20 subjects for a total of 120 images of size 681 \(\times\) 511 pixels. This dataset does not consist of images with human faces, animals or images of activities like sports, music concerts etc. which have semantic meanings attributed to them. The MIT dataset [14] consists of eye fixation records from 15 subjects for 1003 images of size 1024 \(\times\) 768 pixels. Unlike the York University dataset [13], the [14] consists of images which may contain faces, expressions, sporting activities which attract attention and are semantically relevant. The variation in image dimensions, the objects consisted, semantics, topology of arrangement in objects between these two considered eye-gaze datasets offers a test-bed to evaluate the robustness of a saliency map. Also note that the age group of the viewers, the eye-gaze recording equipments, and viewing conditions were not identical while creating these two eye-gaze datasets.

The performance in terms of ROC–AUC\(^{11}\) is measured and the resulting plots for the York University dataset [13] and the MIT dataset [14] are shown in Figs. 2 and 3 respectively. It can be observed from Fig. 2, that the proposed method has a similar ROC plot as compared to graph-based [41] saliency, and clearly outperforms all the other methods considered. The same holds true in case of its performance vis-à-vis the MIT dataset [14] as it can be seen in Fig. 3. To quantitatively evaluate the performance we computed the ROC–AUC which is shown in Table 1. Note that the proposed method has the highest ROC–AUC performance on the York dataset [13] and attains a performance equivalent to the graph-based [41] method on the MIT dataset [14] as it can be observed in Table 1. This is despite the fact that the proposed method is simple and is devoid of the sophistication associated with the graph-based [41], multi-scale [15] or self-information [31] based saliency methods. For the MIT dataset [14] we have not been able to obtain the ROC plots and area under the curve for the self-information [31] and random center-surround approaches [30], since the original source codes of these methods do not work on images as large as 1024 \(\times\) 768 pixels. We preferred not to rescale the input images as this might lead to a bias during the evaluation.

\(^{11}\) Source code used for ROC–AUC computation in this article—http://mark.goadrich.com/programs/AUC/auc.jar.

---

Table 1

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Global contrast [25]</td>
<td>0.54</td>
<td>0.53</td>
</tr>
<tr>
<td>Symmetric contrast [26]</td>
<td>0.64</td>
<td>0.63</td>
</tr>
<tr>
<td>Graph-based [41]</td>
<td>0.84</td>
<td>0.81</td>
</tr>
<tr>
<td>Multi-scale [15]</td>
<td>0.81</td>
<td>0.76</td>
</tr>
<tr>
<td>Entropy-based [29]</td>
<td>0.83</td>
<td>0.77</td>
</tr>
<tr>
<td>Self information [31]</td>
<td>0.67</td>
<td>NA</td>
</tr>
<tr>
<td>Local steering kernel [39]</td>
<td>0.75</td>
<td>0.72</td>
</tr>
<tr>
<td>Weighted contrast [30]</td>
<td>0.75</td>
<td>NA</td>
</tr>
<tr>
<td>Distance transform [28]</td>
<td>0.77</td>
<td>0.75</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.85</td>
<td>0.81</td>
</tr>
</tbody>
</table>
The experimental parameter $n$ is correlated with the scale of the image where the standard setting is $0.02 \times r \times c$. We therefore evaluated the performance of the proposed method on the York dataset [13] by varying $n$ and recorded the ROC–AUC values. It can be observed from Fig. 4 that for 10 random samplings, the proposed method already achieves an ROC–AUC of 0.76 and saturates to an ROC–AUC of 0.85 for 60 random samplings and above. We varied $n$ (where $n=10, 60, 110, 260$ and standard configuration) and obtained ROC plots on the York dataset [13] which are given in Fig. 5. We deliberately included minimal number of ROC plots, as all the ROC plots looked very similar when $n > 60$. Despite this, we recommend to use the standard configuration for $n$. As the size of the image increases so does the possibility that it consists of more objects. In this context, sampling the image in tandem with its size helps in computing a better saliency map. Our method employs random sampling which means that the selected windows at each trail may be different. Hence we decided to study the stability of the proposed method in terms of ROC–AUC for the standard configuration of $n$ on the York University [13] dataset. The results shown in Fig. 6 reveal that the variations in ROC–AUC values over 10 different trails are not significant, hence adding to the stability of the proposed method.

5.2. Experiments on salient region detection task

The MSRA dataset [12] consists of 5000 images annotated by nine users. The annotators were asked to enclose what they thought was the most salient part of the image with a rectangle. Fig. 7 shows an example of this labeling. Naturally occurring objects do not necessarily have a regular contour which can be enclosed accurately inside a rectangle. As it can be seen from Fig. 7, unnecessary background is enclosed by this kind of annotation. It has been recently shown in [25,26,60] that a more precise-to-contour ground truth leads to a more accurate evaluation. Motivated by this a precise-to-contour ground truth was released in [25], for a subset of 1000 images from the MSRA dataset [12]. Thus we consider this subset of 1000 images which have accurate ground truth annotations for our experiments.

In general, the MSRA dataset [12] is significantly different from the eye-gaze datasets in terms of test protocol, content and image size. Fundamentally, the eye movements were not recorded and annotators were required to enclose the most visually interesting region of the image. Such an annotation task involves high-level cognitive mechanisms and is not stimulus driven. Thus there is a weak association between the exact segmentation masks and the saliency of the image. Despite involving a high-level visual task, the MSRA dataset [12] is still relevant to test whether there is an association between the predicted saliency map and the ground truth masks of this dataset. Previous studies [61,57] have shown that the positions of the principal maxima in a saliency map are significantly correlated to the positions of areas that people would choose to put a label indicating a region of interest. In order to quantitatively evaluate the performance for the task of detecting salient regions, we followed the method.

Fig. 4. Variations in ROC–AUC of proposed method due to change in $n$ on York University [13] dataset. Observe that the proposed method has an ROC–AUC value of 0.76 for $n=10$ which is better than the optimal performance of the local steering kernel [39] (ROC–AUC=0.75) and self information [31] (ROC–AUC=0.67) based approaches as seen from Table 1. The method saturates in terms of ROC–AUC for a small value of $n$, and hence rigorous testing to fix this parameter could be avoided. The lack of oscillation in ROC–AUC values versus $n$ shows that the proposed method is stable despite changes in the value of $n$.

Fig. 5. Variations in ROC plots of the proposed method due to change in $n$ on the York University [13] dataset. Observe that for $n=10$, the proposed method obtains comparable ROC plot to local steering kernel [39] and distance transform [28] as seen from Fig. 2. With $n=60$ the performance of the proposed method starts saturating and this helps in processing large images without compromising on run-time.

Fig. 6. ROC–AUC performance of the proposed method for different trails on the York University [13] dataset. Observe that the ROC–AUC values do not change significantly during 10 different trails. The mean ROC–AUC is 0.8554 and the standard deviation ($\sigma$) is 9.3743e–04. The low standard deviation indicates the stability of the proposed method.
not be suitable, as the binarization threshold depends on the weighted contrast based method. However, it could be noted from our previous experiments on eye fixation datasets that the graph-based approach.

The proposed method. Despite being simple, the proposed method fares well in both the tasks of eye-gaze correlation as well as salient region detection and has an equivalent performance to the graph-based method. The resulting recall versus precision curve is shown in Fig. 8. This curve provides a reliable comparison of how well various saliency maps highlight salient regions in images. It can be observed from Fig. 8, that the proposed method has comparable performance with the graph-based method and outperforms all the other approaches in consideration with an exception of the weighted contrast based method. However, it could be noted from our previous experiments on eye fixation datasets that the weighted contrast based method fails to work when the size of the input is large and has far below performance in terms of ROC–AUC as compared to the graph-based approach and the proposed method. Despite being simple, the proposed method fares well in both the tasks of eye-gaze correlation as well as salient region detection and has as an equivalent performance to the graph-based approach.

A common threshold to binarize different saliency maps might not be suitable, as the binarization threshold depends on the image specific statistics. Thus we employ Tsai’s moment preserving algorithm which has also been recommended in to select the map specific binarization threshold. We further evaluate the accuracy of the obtained binary masks to the ground truth masks by employing F-measure as suggested in. The formula for F-measure is given in Eq. (5).

\[ F\text{-measure} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \] (5)

It can be observed from Table 2 that the proposed method and the graph-based method have an identical F-measure performance of 0.64. The weighted contrast-based saliency attains the highest F-measure of 0.71, but does not perform well in the eye-gaze correlation task. The entropy-based method has the lowest F-measure performance of 0.43 despite achieving a good performance on the eye-gaze task. Apart from the proposed method and graph-based saliency, only the multi-scale approach with a F-measure of 0.58 is seen to perform equally well on eye-gaze detection task. We demonstrate the performance of the various approaches considered on the task of salient region detection by considering three sample images from the MSRA dataset as shown in Fig. 9. For the purpose of illustration we overlaid the original image with ground truth masks and the other masks obtained after binarizing the respective saliency maps by Tsai’s moment preserving algorithm. It can be observed from Fig. 9, that the entropy-based saliency approach fails to effectively localize the salient object in the image. In case of the image consisting of the yellow flower, the entire image is shown as salient; while the salient region of the glowing bulb is treated as irrelevant. Despite being efficient the global contrast and symmetric contrast based methods ignore the finer details of the salient region, as both the methods operate on mean pixel intensity value of the image. Conversely, the multi-scale and local steering kernel based saliencies highlight the minute details under consideration.

![Fig. 7. Rectangular and exact segmentation masks.](image-url)

![Fig. 8. Recall–precision curve on the MSRA dataset.](image-url)

<table>
<thead>
<tr>
<th>Saliency map</th>
<th>Recall</th>
<th>Precision</th>
<th>(F)-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global contrast</td>
<td>0.49</td>
<td>0.63</td>
<td>0.51</td>
</tr>
<tr>
<td>Symmetric contrast</td>
<td>0.40</td>
<td>0.66</td>
<td>0.47</td>
</tr>
<tr>
<td>Graph-based</td>
<td>0.73</td>
<td>0.62</td>
<td>0.64</td>
</tr>
<tr>
<td>Multi-scale</td>
<td>0.67</td>
<td>0.56</td>
<td>0.58</td>
</tr>
<tr>
<td>Entropy-based</td>
<td>0.98</td>
<td>0.29</td>
<td>0.43</td>
</tr>
<tr>
<td>Self information</td>
<td>0.55</td>
<td>0.44</td>
<td>0.46</td>
</tr>
<tr>
<td>Local steering kernel</td>
<td>0.48</td>
<td>0.50</td>
<td>0.46</td>
</tr>
<tr>
<td>Weighted contrast</td>
<td>0.84</td>
<td>0.66</td>
<td>0.71</td>
</tr>
<tr>
<td>Distance transform</td>
<td>0.85</td>
<td>0.33</td>
<td>0.45</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.71</td>
<td>0.63</td>
<td>0.64</td>
</tr>
</tbody>
</table>

12 http://www.cs.tut.fi/~ant/histthresh/HistThresh.zip
salient regions as they have a bias towards strong gradients. The self information [31] based approach highlights rare segments of the image as salient, and consequently it discards regions of uniform appearance which can be observed from column nine of Fig. 9. Only the weighted contrast [30] based method has a near perfect performance on detecting the salient regions. But the previous experiments on eye-gaze prediction have shown that the weighted contrast [30] based approach is not highly effective. The proposed method highlights the salient regions along with the finer details effectively which shows that it is not biased towards edges like the other methods.

5.3. Computational run time

We evaluated the run-time of the proposed saliency approach with reference to the other methods in consideration. The run-time of the various methods were benchmarked on three different scales of a color image as shown in Table 3. The original plugins of global contrast [25], symmetric contrast [26], weighted contrast [30], entropy-based [29] and self information-based [31] saliency approaches are pure Matlab codes. While the codes pertaining to local steering kernel [39], multi-scale [15] and graph-based methods [41] are quasi Matlab codes which call C++ functions for run-time optimization. The original plugin for distance transform [28] is a binary executable while its original coding language is unknown. The proposed method is programmed in Matlab. An absolute comparison on the basis of run-time might penalize the methods which are coded in Matlab script as they are relatively slower than their C++ counterparts. Nevertheless, it gives a relative overview of run-time performance of the all methods under consideration.

<table>
<thead>
<tr>
<th>Saliency map</th>
<th>Code type</th>
<th>Run-time (in s) w.r.t. image size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global contrast [25]</td>
<td>Matlab</td>
<td>0.0652</td>
</tr>
<tr>
<td>Symmetric contrast [26]</td>
<td>Matlab</td>
<td>0.0976</td>
</tr>
<tr>
<td>Graph-based [41]</td>
<td>Matlab with C++</td>
<td>0.6256</td>
</tr>
<tr>
<td>Multi-scale [15]</td>
<td>Matlab with C++</td>
<td>0.4388</td>
</tr>
<tr>
<td>Entropy-based [29]</td>
<td>Matlab</td>
<td>2.1448</td>
</tr>
<tr>
<td>Self information [31]</td>
<td>Matlab</td>
<td>1.8466</td>
</tr>
<tr>
<td>Local steering kernel</td>
<td>Matlab</td>
<td>3.0590</td>
</tr>
<tr>
<td>Weighted contrast [30]</td>
<td>Matlab</td>
<td>0.2701</td>
</tr>
<tr>
<td>Distance transform [28]</td>
<td>Binary</td>
<td>0.1266</td>
</tr>
<tr>
<td>Proposed</td>
<td>Matlab</td>
<td>0.3430</td>
</tr>
</tbody>
</table>

Fig. 9. Qualitative analysis of results for MSRA [12] dataset. The first column contains the original image. The remaining columns contain the images overlaid with thresholded (using Tsai’s algorithm [62]) saliency maps obtained from the respective methods mentioned in the figure header. Observe that the proposed method highlights the salient regions along with its finer details effectively, which most of the other methods in consideration fails to achieve.
It can be observed from Table 3 that the run-time of the local steering kernel [39], the graph-based [41] and the multi-scale-based [15] saliency approaches do not change significantly irrespective of the size of the input image. This is on account that the input image is rescaled to pre-specified dimension mentioned in their source codes. The rest of the methods process the input image in its original scale and hence the run-time changes with the input image dimension.

6. Discussion and conclusion

We have compared and contrasted nine existing state-of-the-art approaches to compute a saliency map along with the proposed method for the two different tasks of eye-gaze correlation and salient region detection. Experiments were carried out on large publicly available datasets of York University [13], MIT [14] for eye-gaze correlation and the MSRA dataset [12] for salient region detection. Our results have shown that the proposed method attains state-of-the-art performance on the task of eye-gaze correlation and also has a comparable performance with the other methods for the task of salient region detection. Most of the existing approaches to compute a saliency map fail to perform equally well on both of the aforementioned tasks. But our method along with the graph-based [41] approach has a reliable performance on both these diverse tasks. It should be observed that the proposed method achieves this performance despite being simple as compared to graph-based [41], local steering kernel-based [39] and other existing sophisticated methods. It has also been shown in the experiments that a very small number of random rectangles \((n=60)\) is adequate to attain good results. Despite this we recommend to tune the sampling rate in direct proportion to the size of the image, as generating large number of arbitrary sized image windows often leads to a higher probability of completely enclosing a salient region or an object. In spite of its simplicity and elegance the proposed method fails to capture saliency when color contrast is either low or not being the principal influence. In future, we plan to address this bottleneck and further improve our method to capture higher level semantics like detecting pop-outs in terms of orientation, shape and proximities. Exploiting the ability of the proposed method to detect a salient region as a pre-processor for object detection is also envisaged.

Acknowledgments

Tadmeri Narayan Vikram would like to thank his colleagues Miranda Grahl, Torben Töniges, Christian Schnier and Subhashree Vikram who proof read this article at various stages.

This work has been supported by the RobotDoc Marie Curie Initial Training Network funded by the European Commission under the 7th Framework Programme (Contract no. 235065).

References


Tadmeri Narayan Vikram has received Masters of Computer Applications from University of Mysore, India, in 2006. He worked as a research engineer at GREYC, University of Caen, France between 2008 and 2010. He is awarded the prestigious Marie Curie fellowship since 2010, to pursue Ph.D. at the Applied Informatics Group in Bielefeld University, Germany. His research interests include computational modeling of visual attention and its applications in developmental robotics.

Marko Tscherepanow received his diploma degree in computer science from Ilmenau Technical University, Germany, in 2003. Afterwards, he joined the Applied Informatics Group at Bielefeld University, Germany. In December 2007, he finished his Ph.D. thesis entitled “Image Analysis Methods for Location Proteomics”. His current research interests include incremental on-line learning, neural networks, evolutionary optimization, feature selection, and the application of such methods in cognitive robotics.

Brita Wrede received her masters degree in computational linguistics and the Ph.D. degree (Dr.-Ing.) in computer science from Bielefeld University in 1999 and 2002 respectively. From 1999 to 2002 she pursued a Ph.D. program at Bielefeld University in joint affiliation with the Applied Informatics Group and the graduate program task-oriented communication. After her Ph.D. she received a DAAD PostDoc fellowship at the speech group of the International Computer Science Institute (ICSI) at Berkeley, USA. Currently she heads the Applied Informatics Group at Bielefeld University.