Machine Vision and Applications (2016) 27:893–914
DOI 10.1007/s00138-016-0754-x

ORIGINAL PAPER

Contextual-based top-down saliency feature weighting for target detection

Ibrahim Rahman[1] · Christopher Hollitt[1] · Mengjie Zhang[1]

Received: 25 May 2015 / Revised: 18 November 2015 / Accepted: 12 January 2016 / Published online: 3 March 2016
© Springer-Verlag Berlin Heidelberg 2016

Abstract Target detection using attention models has recently become a major research topic in active vision. One of the major problems in this area of research is how to appropriately weight low-level features to get high quality top-down saliency maps that highlight target objects. Learning of such weights has previously been done using example images having similar feature distributions without considering contextual information. In this paper, we propose a model that we refer to as the top-down contextual weighting (TDCoW) that incorporates high-level knowledge of the gist context of images to apply appropriate weights to the features. The proposed model is tested on four challenging datasets, two for cricket balls, one for bikes and one for person detection. The obtained results show the effectiveness of contextual information for modelling the TD saliency by producing better feature weights than those produced without contextual information.

Keywords Saliency · Detection · Top-down · Features · Contextual · Target

1 Introduction

Many promising human visual attention (HVA) inspired models [16,18,23,43] have been proposed to solve real-world vision problems because of the remarkable ability of humans to perform complex visual tasks efficiently and with high precision. Visual attention represents a a set of cognitive operations that samples the visual field by selecting relevant information and processing them [27]. Two simultaneous mechanisms take place in the HVA system [11]. The first is a fast data-driven mechanism known as the bottom-up (BU) influence which responds to small changes within a visual scene. The BU mechanism directs attention towards those areas that appear markedly different from their surroundings. These regions are called salient regions. The second mechanism, known as top-down (TD), is a slow task-driven mechanism that directs the attention towards targets relevant to the task. The TD mechanism is very important as it is the dominant factor in controlling the gaze and shifting it towards the target/goal during a high-level task [43].

Bottom-up saliency detection or simply saliency detection deals with the detection of objects in an image having salient attributes or features. Most of the literature in visual attention falls into this category. However, BU saliency is not suited for task-driven scenarios such as target detection [11,43] because not every salient object will necessarily be an object of interest.

The majority of the BU saliency techniques are either purely computational-based techniques [1,2,10,15,33,35,39] and several follow the HVA structure or are biologically inspired [16,18]. For TD saliency, most research work is confined to experimental studies [26,36,46,47]. Although some TD saliency computational models have been proposed in the past [4,13,17,28,30,34,45], no unified model describing the TD behaviour exists.

The most successful and widely used approach to modelling the TD saliency is by introducing weights to the low-level BU features [4,13,28,30,34]. A main challenge in feature weighting is how dynamically the weights are assigned to the features. Previously, while different weight learning approaches have been adopted, most of these approaches suffer from being static. This means that the learnt weights can only work in similar types of examples,

✉ Ibrahim Rahman
hmmiharbi@yahoo.com; Ibrahim.rahman@ecs.vuw.nz

[1] School of Engineering and Computer Science, Victoria University of Wellington, Wellington, New Zealand

in this case images, and fail when the examples/images are different in terms of content (e.g., objects, background, etc.).

Image content (or context) is a broad term that describes high-level information within an image such as background, distractor, target, semantic and other prior knowledge. It has been shown experimentally that the inclusion of contextual information improves the efficiency and accuracy of discriminating the target object from the background distractors [36,38,41], though the context idea has not been applied in TD saliency and feature weighting.

To dynamically model the TD saliency, this paper introduces a mechanism that utilizes the contextual information of an image to dynamically assign appropriate weights to the BU features. The inclusion of contextual information for BU feature weighting has not been considered in previous works. Hence, such dynamic feature weighting constitutes the main objective of this research work. We use target detection as an example application that allows demonstration of our approach. Our results show the importance of the contextual information for dynamic TD saliency feature weighting.

The remainder of the paper is organized as follows. After the related work is given in Sects. 2, 3 introduces the major components and the working of the proposed top-down weighting technique. More details about the bottom-up saliency features and the saliency generation itself will be covered in Sect. 4. The contextual information extraction and multiple weight generation will be covered in Sect. 5. A thorough discussion on the results will be presented in Sect. 6. Finally conclusions will be discussed in Sect. 7.

2 Related work

2.1 Bottom-up saliency models

For the last two decades, BU saliency techniques have been intensively used for salient object detection [22,24,32,40,44], segmentation [1,9,25] and recognition [12,28,37]. The performance of these techniques is evaluated in terms of their accuracy in detecting the salient/target objects [21,40] or in terms of their efficiency in computing saliency maps [1,2]. Furthermore, based on the ability of these techniques to predict region of interests, they are classified into either being used for fixation detection [14,20,35,39] or salient object detection [9,21,22].

It would be impossible to discuss all the BU saliency and fixation techniques in this paper; however, a detailed comparison and description of the most recently proposed techniques and models are given in [6,7]. Briefly, to highlight some of the BU state-of-the-art existing saliency techniques, we split these techniques into classical and recent ones.

In classical techniques, for instance in [1] the authors developed a fast technique that works in the frequency domain. This technique is well known for its segmentation performance as boundaries are retained in the generated saliency maps. In [15], authors used a graph-based technique to perform the normalization of the feature and conspicuity maps acquired from Itti model. In [35] statistical spatial rarity of image pixels is exploited to predict eye fixation regions. On the other hand, analysing global and local regions in an image greatly captured the interest of many for saliency detection [10,39]. In [39], patches are extracted from the original images and dissimilarity measures consisting of centre, spatial and colour distances are used to distinguish a salient region from a non-salient one. Similarly in [33], the authors proposed a surprise element called information divergence measure IDM that computes the information variation between image patches in a global sense. The IDM is computed only over the principal component analysis (PCA) colour contrast feature of the patches. Authors show that this model outperforms most of the classical saliency detection techniques on benchmark datasets.

In the last 3 years, saliency detection techniques have become more efficient while attaining a very high accuracy on benchmark saliency datasets. As few examples from the recent development on saliency detection, in [32], a propagation method is proposed that exploits the intrinsic relevance of similar cells or regions based on the information contained in neighbour cells. The authors further proposed a Bayesian framework for multiple saliency map integration. In another similar work [24], the authors proposed a double saliency propagation technique which is based on applying low-level boundary cues for background detection and high-level objectness cue for foreground object detection. The novelty of the technique is highlighted by the fact that the most certain object and boundary superpixels are able to propagate saliency knowledge to leverage their complementary influence. The use of high-level cues is a common practice in most of the state-of-the-art saliency detection techniques. For instance, the authors extracted three high-level cues, namely, objectness, uniqueness and focusness and fused them together to determine the saliency of regions [19].

Another set of methods concentrate on grouping foreground or background pixels/superpixels having similar attributes or features. For instance in [22], regularized random walks are used to establish pixel-wised saliency maps from the background and foreground superpixels whereas in [44], a graph-based manifold ranking approach is adopted.

Some state-of-the-art techniques still follow the classical colour contrast approaches for saliency detection. This is based on the human perception observation that salient regions often have distinctive colours compared to the background. In [21], the authors show that salient regions can be linearly separated from the background in high-dimensional colour space. In another contrast approach [9], the authors introduced an efficient histogram-based contrast method

(HC) to measure saliency. Furthermore, they incorporated spatial regional factors to HC method to produce a more effective saliency measure called regional-based contrast (RC).

Other methods include the use of supervised machine learning and classifiers approaches directly on positive and negative image samples to learn a strong boosting classifier from a set of weak classifiers [40], the use of free energy principles that computes the entropy between input image a reconstructed copy of the same image [14], decomposing images into abstract representation to remove unnecessary image details to allow more effective saliency assignemnt to salient pixels [8], and the incorporation of quantum mechanics into graph-cut for accurate saliency region segmentation [3].

Most of the state-of-the-art saliency techniques have achieved very high accuracy performance in most benchmark saliency datasets. However, in some other challenging datasets where saliency attributes are more challenging (e.g., out of focus, no centre-bias factor, complex background, low contrast, etc.), the performance of these techniques degrades [7]. Furthermore, the implementation of some of these techniques for real-time applications is not suitable due to low efficiency. Mostly this is due to the use of high-level features and complex feature manipulation procedures.

2.2 Top-down saliency models

Perhaps one of the most prominent HVA-based saliency models is that of Itti et al. [18] based on the famous integration theory of Treisman and Gelade [42]. The majority of the extant TD feature weighting mechanisms use the Itti attention model for feature extraction and integration. Our model also follows the structure of the Itti model. As shown in Fig. 1, the process of Itti model begins with the extraction of low-level features from three channels, colour, intensity and orientation at various scales. A centre-surround mechanism is applied on the features that uses the Difference of Gaussian filters to produce the feature maps (FM). The channel wise FMs are then integrated and normalized at different scales to yield the conspicuity maps (CM). Finally, the CMs are combined linearly to build the final saliency map. The saliency map generation is followed by a Winner-Take-All (WTA) and Inhibition of return (IOR) mechanisms for fixation and gaze shift, respectively.

This model is used to generate the BU saliency maps for salient object detection. However, a proper weighting of the FMs and CMs as indicated by many authors results in generation of a so-called top-down saliency maps tuned for a particular task [4,13,28,30,34].

The quality/accuracy of such weights depends on how they are learnt or estimated. In [13,28], the authors proposed a system called VOCUS that learns the most salient regions (MSR) using a classifier. The weights of individual features are calculated as the ratio of the average saliency of the MSR regions and rest of the background region.

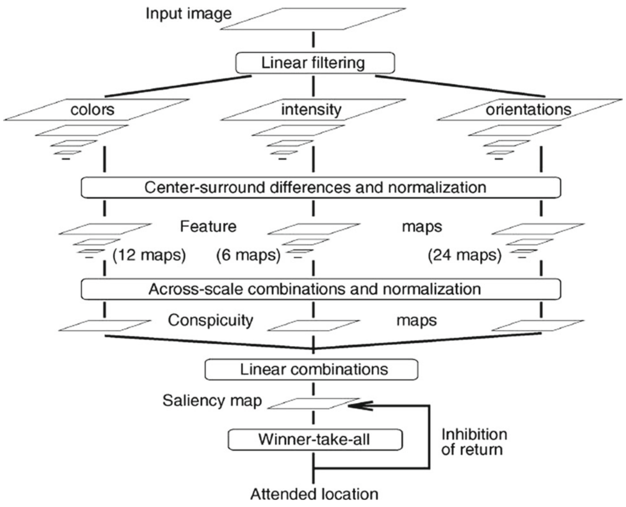


Fig. 1 Itti attention model [18]

In another work by Benicasa et al. [4], a recognition indicator is used to find the likelihood of a segmented region belonging to a salient object. This is achieved by appropriately adjusting the weights through a feedback loop from a high-level classifier.

In [20], the authors learned feature weights from human observed fixation data through linear support vector machine (SVM) classifier. The learned weights from the classifier are used to weight various features to model and predict where human usually look when performing a free viewing of natural scenes. They used BU Itti features, mid-level horizon cues and high-level person and face detectors. The proposed model accurately predicted fixation regions compared to the groundtruth human observed fixations.

Probably the most prominent TD weighting model is the one proposed by Navalpakkam and Itti [30]. The weight calculation process is formulated as an optimization problem by maximizing the signal to noise ratio (SNR) such that the signal and noise represent the average saliency energy of the target and that of the distracting background, respectively. The weights are calculated within feature dimension indicated by $g_{i,j}$ (i.e., sub-feature j of a feature channel i) and across features g_j (i.e., CMs) where j is the CM. These optimized weights are given as

$$\begin{aligned} g_{i,j} &= \frac{\mathrm{SNR}_{ij}}{\frac{1}{n}\sum_{k=1}^{n}\mathrm{SNR}_{kj}} \\ g_j &= \frac{\mathrm{SNR}_j}{\frac{1}{N}\sum_{k=1}^{N}\mathrm{SNR}_k} \end{aligned} \tag{1}$$

where n is the number of sub-features within a feature channel and N is the number of CMs.

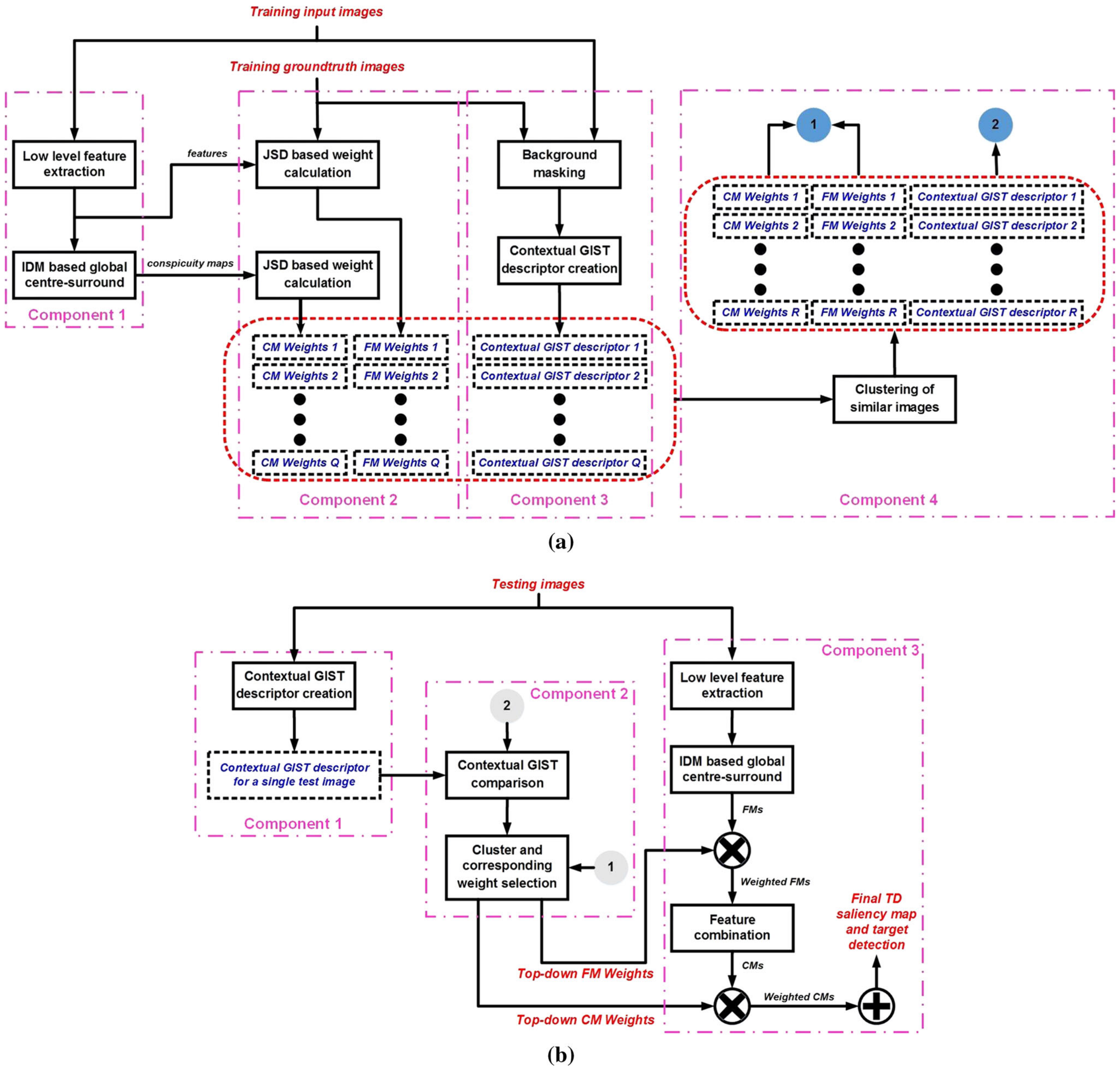
A common problem in all these approaches is that they do not consider any high-level information when evaluating the feature weights. Features that are learnt or evaluated in this way are static and achieve good performance on similar types of images/examples in terms of context (i.e., similar images in both the training and testing phases) but perform poorly when there is variation in image context. To make the feature weighting a dynamic process, we propose a model that incorporates high-level contextual information of the images to dynamically assign weights to the features and can be applied on a variety of images with different context.

The main objective of this paper is to use image contextual information to assign dynamic and appropriate weights to the BU features. To implement this concept, three sub-tasks are performed. Furthermore, besides the collective contribution of the sub-tasks in building the final model, each sub-task can be considered to be a separate model. The sub-tasks and their respective contributions towards building the final model are as follows,

1. *To have a set of good BU features and an effective saliency map generation method*
Task: To build the final TD model, the first step is to have a set of good features to perform the weighting on. In addition, a good saliency map generation method from the features plays an important role in building an effective attentional-based vision model. In the past two decades, several saliency detection techniques were proposed that utilized different features and methods to produce quality saliency maps. Our proposed saliency map generation method is inspired by Itti's feature model and by the centre-surround mechanism proposed in [33].
The reason for choosing Itti features is because they are computationally efficient than other complex high-level features utilized by other saliency models. In addition, mid-level features are added to the basic ones. This is done for two reasons, (1) To have a richer set of features for a quality saliency map generation, and (2) To have a larger set of features to work with to demonstrate our proposed feature weighting mechanism. Similarly, we used the information-based centre-surround mechanism proposed in [33] as it is fast and achieved a very good performance in saliency detection comparable to state-of-the-art techniques.
Contribution: By performing this task, an effective and efficient saliency detection technique is established. This technique is used as a framework for the contextual-based dynamic feature weighting model. In addition, the technique can be used solely for BU salient object detection (i.e., without any feature weighting). Hence, as one of the sub-contributions of this paper, an accurate BU saliency detection technique is produced having comparable performance to state-of-the-art BU salient object detection techniques.
2. *To calculate the feature weights using the Jensen–Shanon Divergence (JSD)*
Task: Weight calculation is an essential part of the feature weighting model. Previously the SNR approach was used for weight calculation [30]. However, SNR calculation results are unbounded weight values. In addition, SNR is calculated by finding the ratio between the mean pixel intensity values of the target region and that of the background region. This could lead to inverted weight assignment to a feature map in situations when the target region has lower intensity values than the background. A better approach is to use the difference between the distributions of both regions. For this reason, JSD which is a bounded distribution-based measure is used instead.
Contribution: Upon performing this task, a more accurate feature weight computation method is established. This weight computation techniques is used when learning the feature weights by the contextual dynamic feature weighting model. Hence, as another sub-contribution of the paper, an accurate feature weight calculation procedure is established.
3. *To cluster images into groups based on their contextual information during the training phase*
Task: A contextual descriptor is generated for each image and then similar contextual images are clustered during the training phase. Distinct weights are computed for each cluster using the previous mentioned weight computation procedure.
Contribution: Upon performing the task, a new contextual-based image clustering technique is produced. The contribution of this part of the model is the generation of multiple set of possible weights to be assigned to a new test image depending on its context.

Upon completing the above-mentioned tasks, the following contributions are made from this paper,

1. Main contribution 1: A TD model is developed that utilizes contextual information of an image for dynamic assignment of weights to the BU features which was lacking in previous saliency models.
2. Main contribution 2: We show that contextual-based TD weighting outperforms both TD weighting without context and pure bottom-up (unified weighting) approaches.
3. Main contribution 3: We demonstrate that the proposed contextual-based model has better or comparable accuracy performance than various state-of-the-art saliency detection techniques for target object detection in four challenging datasets.
4. Sub-contribution 1: A new developed BU saliency technique based on low-level features and information diver-


Training input images
Training groundtruth images
Low level feature extraction
features
conspicuity maps
IDM based global centre-surround
Component 1
JSD based weight calculation
JSD based weight calculation
CM Weights 1
CM Weights 2
CM Weights Q
FM Weights 1
FM Weights 2
FM Weights Q
Component 2
Background masking
Contextual GIST descriptor creation
Contextual GIST descriptor 1
Contextual GIST descriptor 2
Contextual GIST descriptor Q
Component 3
1
2
CM Weights 1
FM Weights 1
Contextual GIST descriptor 1
CM Weights 2
FM Weights 2
Contextual GIST descriptor 2
CM Weights R
FM Weights R
Contextual GIST descriptor R
Clustering of similar images
Component 4
Testing images
Contextual GIST descriptor creation
Contextual GIST descriptor for a single test image
Component 1
Component 2
2
Contextual GIST comparison
Cluster and corresponding weight selection
1
Top-down FM Weights
Top-down CM Weights
Component 3
Low level feature extraction
IDM based global centre-surround
FMs
Weighted FMs
Feature combination
CMs
Weighted CMs
Final TD saliency map and target detection

(a)

(b)

Fig. 2 The proposed TDCoW model: **a** training phase, **b** Testing phase

gence mechanism for salient object detection that has good detection performance compared to state-of-the-art BU saliency techniques.

5. Sub-contribution 2: A new feature weight computational technique based on JSD is developed which is more effective than the previously proposed SNR mechanism.

3 The proposed model: top-down contextual weighting (TDCoW)

In the proposed model the contextual information represents only the gist of the image. Hence, throughout this paper, gist and context are used interchangeably. The proposed model, which we refer to as top-down contextual weighting (TDCoW), is divided into two phases, the training and testing phases. As shown in Fig. 2a, the first step of the training phase involves feature extraction and the generation of the FM, CM and BU saliency map (SM) using a centre-surround mechanism called information-divergence measure (IDM). This is done for each of the Q training images.

In the second step, the computation of the feature weights for each training image takes place. The weights are calculated by finding the Jensen–Shanon divergence (JSD) of the target object with respect to the background. The weights are calculated for each FM and CM separately.

The third step in the training phase involves contextual information extraction. This is performed by initially masking the target from the background using the ground truth images, so that the context is only extracted from the background region. This is followed by the creation of the feature descriptor from the masked region. The descriptor acts as a context identifier for the image which can later be used for contextual matching. Hence, each training image is associated with FM weights, CM weights and a contextual descriptor as shown in Fig. 2a.

In the fourth step of the training phase, grouping of the training images into R clusters is performed. The grouping is done according to the contextual similarity amongst images using unsupervised k-mean clustering. For each cluster, a contextual descriptor is calculated as the average of the individual contextual descriptors of the images belonging to that cluster. Similarly, the weights of the individual images in a cluster are averaged to yield consolidated weights for that cluster.

The second phase of the model is the testing phase in which a target object is detected. As shown in Fig. 2b, the process begins by creating a contextual descriptor for the test image as in the training phase. However, now the context must be calculated over the whole image without any target masking due to unavailability of the ground truth. There will be some perturbation in the contextual descriptor but we anticipate this to be minor as the ratio of target region to the background is small. This contextual descriptor is compared with the centroid descriptor of each learned cluster using an appropriate an information theoretic distance metric. The cluster with the lowest distance corresponds to the best match for the test image. Accordingly, the matched cluster's weights are selected to be used as the TD weights for the test image. In this way, appropriate weights are assigned to the FMs and CMs of the test image.

Note that multiple set of weights and contexts is learnt during the training phase, one for each cluster. The test phase acts as context template matching process that results in a best possible weight assignment to the FMs and CMs of the test image. The dynamic nature of the model comes from the fact that different weights can be assigned to the test image according to its context as well as that of the learned clusters. A more detailed description of each step in the training and testing phase is described next.

4 Bottom-up feature extraction

The initial step in generating the saliency maps is feature extraction. Eight features are extracted as shown in Fig. 3. These features are colour (C), intensity (I), orientation (O), contrast (Co), centre-bias (Cb), principal component analysis features (PCA), edges (Ed) and frequency-based features (MSS) with each feature category having sub-features. The features range from low-level such as colour and intensity to higher level features such as edges and PCA features. There are two reasons to have many extracted features. First some high-level features such PCA, edges and frequency are assumed to give insight to valuable information about the structure and behaviour of the image. Such information can be very useful for better saliency estimation and and target detection. Secondly, some targets may require additional features or different subsets of features in order to detect them. Itti's basic features might not be sufficient for target detection. On the other hand, there is no specific number or type of features that are assumed to be sufficient for a general target detection. As a reasonable set of features for target detection, the above-mentioned features are used which is a combination of low, high and efficiently computed features.

The colour feature has four sub-features; red (r), green (g), blue (b) and the quantized colour feature (q). The quantization is performed as follows [9]: assuming an input colour RGB image Im is of size $H \times W \times 3$ with 8 bits colour depth, the first step is to quantize the colour range by selecting 12 uniformly distributed levels for each colour channel. This will yield 1728 different possible colours. Furthermore, only 5 % of the most occurring colours in natural images are retained. This is done by observing the most frequent colours from a large database of natural images. The images are quantized with these levels yielding a palette of 85 colours. The quantization is done to reduce the histogram space from 256^3 to only 85 for the subsequent IDM calculation.

The intensity feature has a single sub-feature denoted as I and the orientation features are extracted at 0°, 45°, 90° using Gabor filters. Contrast sub-features are also computed as they have good performance in saliency detection [9,10,39] which includes the Red-Green (rg), Blue-Yellow (by) and Hue (h). These sub-features are given as

$$\begin{aligned} rg &= \frac{r-g}{\max(r,g,b)} \\ by &= \frac{b-\min(r,g)}{\max(r,g,b)} \\ h &= \frac{180}{\pi}\arctan\left(\frac{\sqrt{3}(g-b)}{2r-g-b}\right) \end{aligned} \tag{2}$$

It is worth mentioning here that the colour features are object level features and more suited for object detection whereas colour contrast is more effective for saliency detection. This is the reason for including both types of features.

The next feature is the PCA-based features. PCA is a statistical approach that transforms a set of correlated observations/features into orthogonal uncorrelated segments called principal components (PCs). Irrelevant details and noise are neglected while finding these PCs. The obtained

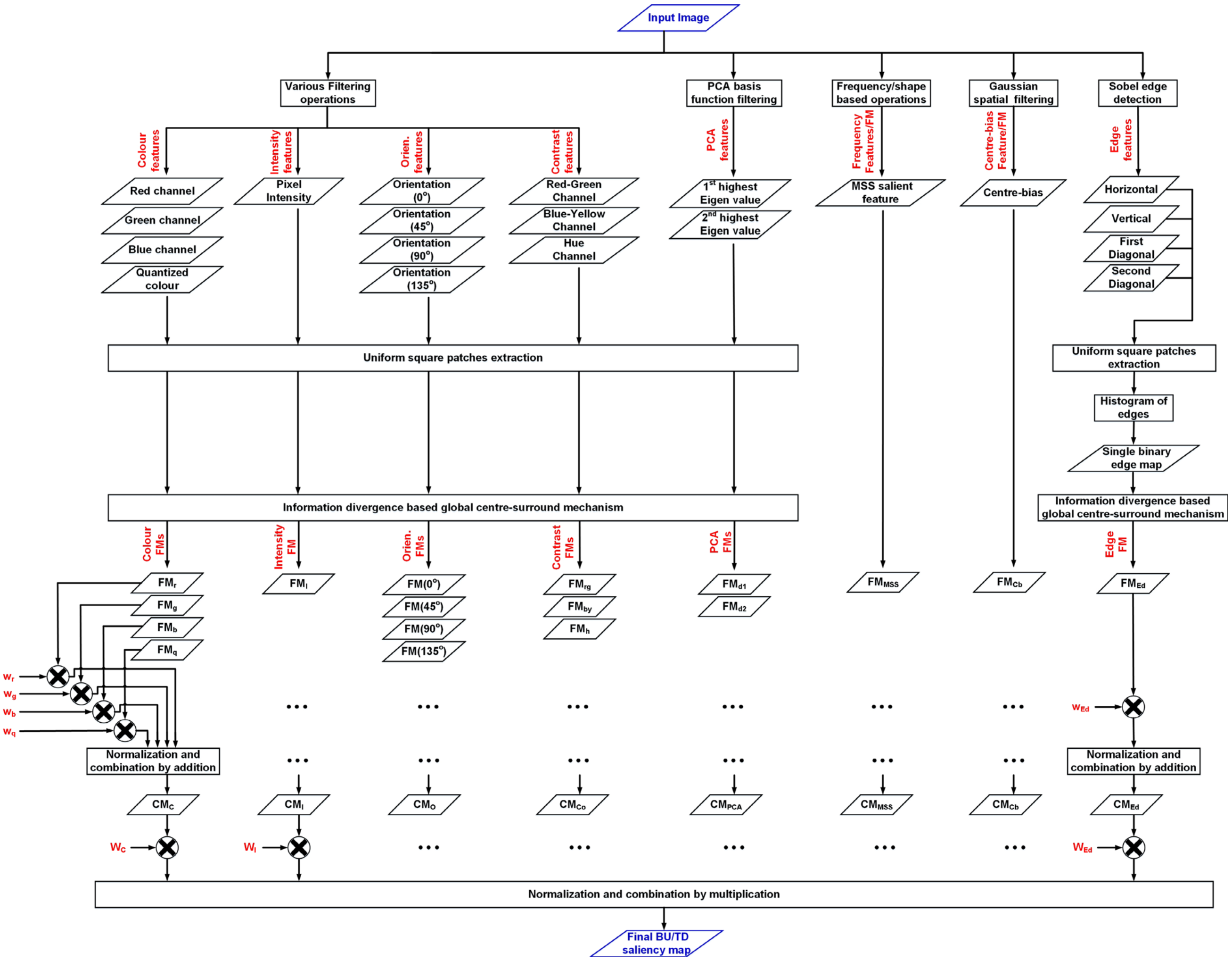


Fig. 3 Feature extraction and saliency map generation procedure

PCs are assumed to describe important features contained in the data, in this case images. Previously, PCA was used for extracting useful features for salient object detection [10,33,45]. One way to implement PCA on the input image is to consider square patches of the input image and then to extract the PCs of the patches separately [33]. However, in this work and for better efficiency, PCA is applied colour wise rather than patch wise as the number of patches $L \gg 3$. We only have three colour channels in the latter to apply PCA on, making this approach more efficient.

The input colour image '*Im*' of size $H \times W \times 3$ is reshaped in such a manner that each colour layer is transformed into a single dimension column vector of length $H \times W$ denoted as $\mathbf{x}_i$. After concatenating the three layers column wise, we get a reshaped layer matrix $\mathbf{X}$ representation of the image such that $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3]$.

The mean value of each column in $\mathbf{X}$ is subtracted from the corresponding column. This is followed by the computation of the covariance matrix of $\mathbf{X}$ and then the eigenvectors. Each eigenvector represents one of the PCs of a total of three PCs denoted as d_1, d_2 and d_3. These components are ordered according to the magnitude of their eigenvalues such that d_1 is associated with the highest eigenvalue. In our experiments, it has been observed empirically that only the first two components prove useful.

The next feature is the edge map which is extracted in vertical, horizontal and diagonal directions. Although there are four sub-features for the edge feature, these are considered as a single sub-feature denoted as '*Ed*' for a reason mentioned later in this section.

A centre-bias factor is added as an additional feature which is represented by a 2-D Gaussian function centred at $(\frac{W}{2}, \frac{H}{2})$ and controlled by the standard deviation value of the function. The centre-bias behaviour models human prior knowledge that the centre of photographs tends to be more salient or contain the target. However, in the test examples, such factor exhibits low weight value which suggests that the target in most of the test images is positioned randomly in an image and not in the centre of the image as it is the case in most of the saliency datasets. The inclusion of this factor is only to show that the datasets chosen for testing our model are competitive for target detection.

The final feature denoted as MSS highlights the frequency distribution of the pixels. This frequency-based feature extraction method is proposed by Achanta [1] for salient object detection and has good performance in salient object segmentation. It exploits all the low frequency components and majority of the high frequencies and considers the position and scale of the objects.

Once the features are extracted, the FMs are generated by calculating the IDM of various patches of the image as proposed in [33]. Information-divergence measure is a centre-surround mechanism that exploits the element of surprise by finding the divergence of information between various regions of the image. Hence, it is responsible for generating the FMs that highlight the possible salient regions within a feature.

This procedure is the same for all the sub-features except for the edge and the centre-bias. Briefly, the process starts by dividing a feature image into smaller regions by uniformly segmenting it into square non-overlapping patches of size $n \times n$. The total number of complete patches is $L = \lfloor H/n \rfloor \times \lfloor W/n \rfloor$. For instance, if an image has the dimension of 12×12, and if the patch size is 5×5, then we will have only $L = 4$ complete patches which are indexed as $i = 1, 2, 3, 4$. We denote these patches by $p_i^j(k)$ where $i = 1, 2, \ldots, L$ represent the patch index, k is the feature notation for which the FM is to be generated and j is the sub-feature notation for the respective feature. For instance $p_1^r(C)$ is the first patch for the sub-feature red belonging to the colour feature.

The IDM is calculated for each patch by finding the divergence of the distributions between two patches for a particular sub-feature. The first patch (centre patch) is one of the patches from $p_i^j(k)$ where as the second patch (surround patch) is the collection of the remainder of the patches as illustrated in Fig. 4. If G and S represent the kernel density estimated (KDE) distributions for the centre and surround regions, respectively, a patch saliency is found by

$$\mathrm{IDM}(i, j, k) = \sum G_i^j(k) \log \left(\frac{G_i^j(k)}{S_i^j(k)} \right) \tag{3}$$

For the centre-bias, the FM is the feature image itself and there is no need to evaluate the IDM. In addition, for the edges, the IDM is calculated in the same way but not directly on the sub-feature edges images. Initially a histogram of edge orientation is evaluated for each patch from the four sub-feature binary edge images. This is followed by IDM calculation but now on discrete edge histograms.

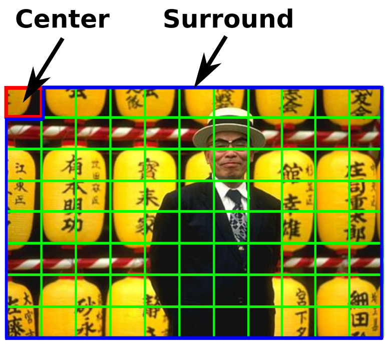


Fig. 4 Global centre-surround patches. The patch within the *red square* represents an active centre patch selected amongst the *green squares*. The collective surround patch is highlighted by the *blue region*

The CMs are generated by summing the weighted sub-features within a feature. Finally the weighted CMs are combined by multiplying them together to yield the final saliency map. Multiplication of CM is better than adding them when combining different types of features as the maps tend to have different spatial distributions and thus we want to have common regions of interest from each map. Map integration procedure can be summarized by the following equation

$$\begin{aligned} \mathrm{SM} &= \mathcal{N}\left(\prod_{k=1}^{8} W_{\mathbf{u}(k)} \left(\sum_{j=1}^{\mathbf{v}(k)} w_j^{\mathbf{u}(k)} \left(\mathrm{FM}_j^{\mathbf{u}(k)} \right) \right) \right) \\ \mathbf{u} &= (C, I, O, \mathrm{Co}, \mathrm{PCA}, \mathrm{Cb}, \mathrm{Ed}, \mathrm{MSS}) \\ \mathbf{v} &= (4, 1, 4, 3, 2, 1, 1, 1) \end{aligned} \tag{4}$$

where $w_j^{\mathbf{u}(k)}$ is the weight associated with each FM and $W_{\mathbf{u}(k)}$ is the CM weight. The tuples $\mathbf{u}$ and $\mathbf{v}$ represent the feature symbol tuple and the number of sub-feature tuple corresponding to each feature, respectively. Note that $\mathcal{N}$ indicates a normalization step before acquiring the final saliency map to promote very strong peaks and suppress the rest. This step is essential to obtain a saliency map with only the most dominant salient region being highlighted.

5 Contextual feature representation and feature weighting

This section describes how contextual information is extracted and then used to assign dynamic weights to the low-level features. In addition, an effective weight calculation procedure is explained that depends on the JSD between a target and rest of the image.

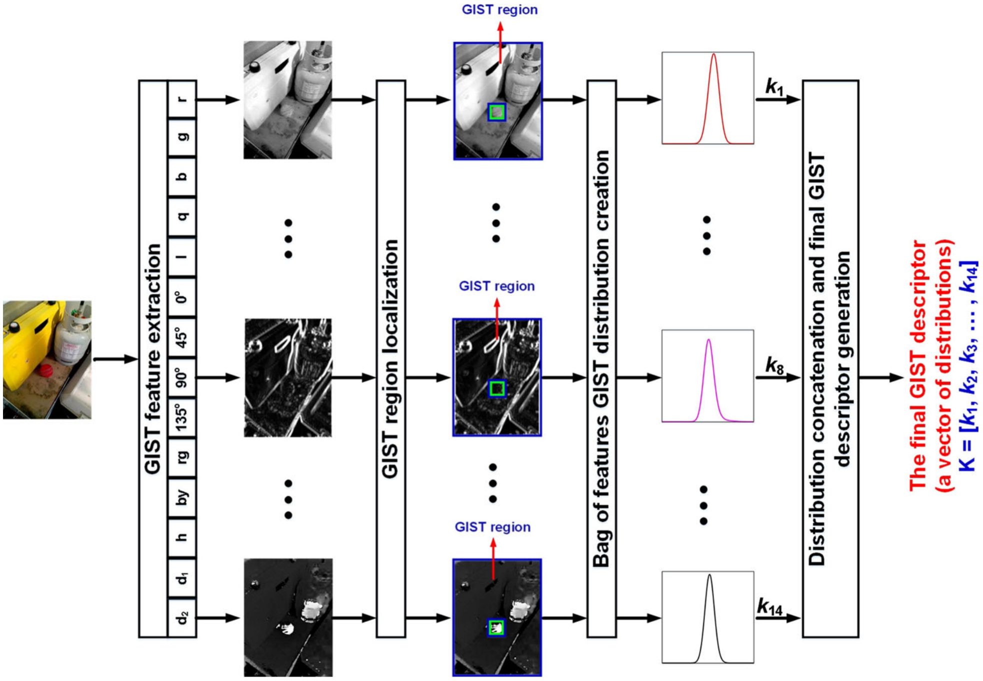

GIST region
GIST region
GIST region
GIST feature extraction
r
g
b
q
I
0°
45°
90°
135°
rg
by
h
d1
d2
GIST region localization
Bag of features GIST distribution creation
k_1
k_8
k_{14}
Distribution concatenation and final GIST descriptor generation
The final GIST descriptor
(a vector of distributions)
K = [k_1, k_2, k_3, … , k_{14}]

Fig. 5 An example of contextual descriptor construction. From *left* to *right*: input image, extracted features, masking regions (*green region* is excluded and the *blue region* is the gist region for which a descriptor is created), distributions of the gist regions, concatenation of the distributions and the final contextual descriptor vector

5.1 Contextual descriptors

A large contextual descriptor of distributions for an image is created using colour, intensity, orientation, contrast and PCA features. Such image descriptors are sometimes called bag of features and are commonly used in classification problems [5,48]. The distribution is estimated as before using KDE for a fixed number of sample points. Hence, the size of the descriptor depends on the KDE number of estimation points. Large number of sample points corresponds to large descriptors that will impose high computational demands for contextual matching. At the same time, we want to avoid under-sampling which can produce inaccurate results. Empirically we set the number of sample points to 1000. Figure 5 shows the complete procedure for generating the contextual gist descriptor.

5.2 Feature weighting

Previously in [30], weights were evaluated by calculating the signal (the target) to noise (distractors) ratio from the FMs. In this paper, a more effective weighting calculation procedure is proposed. There are two shortcomings of the SNR weighting mechanism. First, the weighting information is extracted from the sub-features directly and not from the FMs. The value of the weight is highly dependent on how the FMs are generated. For instance for a particular sub-feature, if the corresponding FM is generated incorrectly (the target region is less highlighted than the background region), then according to the SNR Eq. (1), such an FM will receive a low weight whereas it should be assigned a high value. Hence, a poor FM generation algorithm could result in a wrong weight assignment to the feature when SNR approach is used.

Secondly, SNR is calculated by finding the ratio between the mean pixel intensity values of the target region and that of the background region. Again this could lead to an inverted weight assignment to a CM in situations when the target region has lower intensity values than the background within the CM. This is because the ratio/difference is taken between two single values (i.e., the mean values of the two regions).

To overcome the problems associated with the SNR-based approach, a distribution-based weight calculation approach is adopted using JSD. Jensen–Shanon Divergence is a bounded dual version of the IDM. The divergence between the target region having a distribution h_T and the background region with distribution h_B is calculated directly from the sub-features rather than the FM to avoid the involvement of the FM generation algorithm. The weight is calculated as follows

$$\mathrm{JSD}\,(T||B) = \frac{1}{2}\left(Z_{\mathrm{IDM}}\,(T||M) + Z_{\mathrm{IDM}}\,(B||M)\right)$$

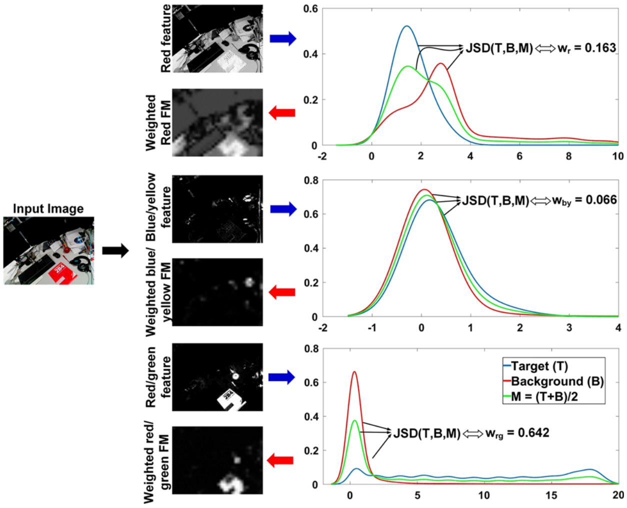


Fig. 6 JSD-based top-down weight calculation example for three sub-features, *red*, *blue/yellow* and *red/green* as can be seen left to the *blue arrows*. Distributions of target (T), background (B) and the intermediate region (M) for the respective sub-features are shown on the *right column*. According to the variation in distribution, a weight value is calculated using (5) as 0.163, 0.066 and 0.642, respectively, for the three sub-features. The images on the left of the *red arrow* show the impact of the weights on the FMs. For this example, the best map in terms of detecting the target *red ball* is the *red/green* feature which is assigned the highest weight value compared to the other sub-features

$$h_M = \frac{1}{2}(h_T + h_B)$$

$$Z_{\mathrm{IDM}}\,(T||M) = h_T \log_2\left(\frac{h_T}{h_M}\right)$$

$$Z_{\mathrm{IDM}}\,(B||M) = h_B \log_2\left(\frac{h_B}{h_M}\right) \qquad (5)$$

where h_M is the histogram of the intermediate region M. The terms $Z_{\mathrm{IDM}}\,(T||M)$ and $Z_{\mathrm{IDM}}\,(B||M)$ are the target/intermediate and the background/intermediate regions' IDMs, respectively.

It is worth mentioning here that a statistically high JSD value for a particular sub-feature suggests that the target region is highly different from other regions of the image. This indicates the importance of that sub-feature for the target detection and should be assigned a high weight value. As an example, Fig. 6 shows how the weights vary due to the statistical information difference of the target and the background for some sub-features and this can be reflected in the corresponding weighted FMs. For instance, the distributions of the target, background and the intermediate region M for the red/green sub-feature have high variation. Hence, the assigned weight to this sub-feature is relatively high (i.e., 0.64). This can be observed by the corresponding FM which highlights the target red ball better than the other two sub-features.

5.3 Clustering and contextual matching

The main objective of TDCoW is to allocate the test image appropriate weights according to the similarity of its contextual contents with that of the training images. This is done by matching the contextual descriptor of the test image with the training images. However, increasing the number of training images increases the processing time for matching. As a result, a reasonable choice would be to cluster similar images together to reduce the search space for matching according to the contextual similarity.

The clustering approach is similar to the classical k-mean clustering except for the distance measure calculation. The clustering is done on the contextual descriptor of the training images over several iterations. After every iteration, clusters are created by measuring the JSD between descriptors. The reason for using JSD and not the traditional Euclidean distance is that we are measuring distance between distributions. Furthermore, once clusters are created after a single iteration, a mean descriptor for that cluster (called a centroid descriptor) is re-calculated from the individual contextual descriptors of the images belonging to their respective clusters. This is achieved by considering the joint distribution of the individual descriptors of the cluster.

After the final iteration, the centroid of each cluster represents the final contextual descriptor for the corresponding cluster. Furthermore, the learned weights of the individual images of a cluster are averaged to yield a single set of weights for that cluster. Hence, at the end of the clustering process, R number of descriptors and weights are produced, one for each cluster.

When a test image is given, the ultimate objective is to assign appropriate weights to the features which are learned from the training phase to produce the final TD saliency map. The following steps are followed,

1. Creating a contextual descriptor for the test image by following the same procedure applied to a training image which is explained in Sect. 5.1.
2. The contextual descriptor of a test image is matched with the contextual descriptor of each of the clusters being created in the training phase. The distance computation to perform the matching is accomplished using JSD. Note that JSD here is used for another purpose. Previously in the training phase we have used JSD to calculate the weight of FM and CM based on finding the target and background distributions. However, here JSD is used to see the information distance between the contextual descriptor of the test image and that of each cluster. Note that the same JSD Eq. (5) is used. The only difference is that the target distribution is replaced by the contextual

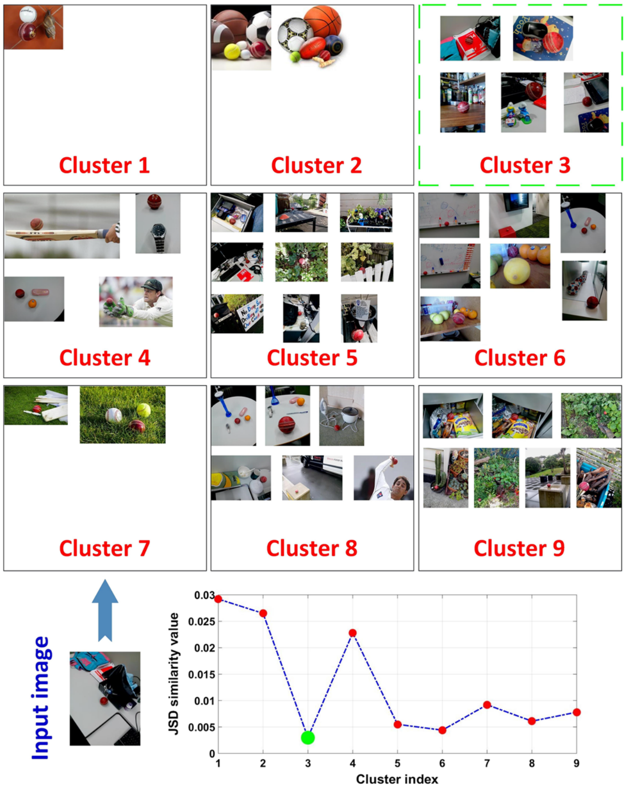


Fig. 7 Contextual-based image clustering illustration. 43 images are clustered into 9 groups using the proposed k-mean contextual clustering technique. The similarity of images within a cluster in terms of the used contextual features can be seen in some clusters e.g. cluster 4 and 5. An input test image is contextually matched to these clusters using JSD with cluster 3 being the best match [lowest JSD value (see the *bottom plot*)]

distribution of the test image and the background distribution is replaced by the contextual distribution of one of the clusters for which the matching is taking place.

3. The cluster having the lowest JSD value corresponds to the best match to the input image.
4. The precomputed weights of the FM and CM of the best match cluster are assigned to the input image.
5. The final TD saliency map using the selected weights is computed to produce the final result by following the saliency map generation procedure described in Sect. 4.

Following these steps, a dynamic weighting of BU features is achieved based on the contextual content of the image.

Figure 7 shows a sample of 43 training images being grouped into nine clusters by the above method. Furthermore, a sample test image is matched with the centroids of each cluster using JSD.

6 Results and analysis

The experiments are divided into three parts. In the first part, we show the effectiveness of the proposed features and the global IDM centre-surround mechanism over the traditional features and centre-surround mechanism proposed in the Itti model. This evaluation is further extended for salient object detection i.e., BU saliency and compared with some state-of-the-art techniques in saliency detection. The second part of the experiment deals with the nature and effectiveness of the weighting calculation method. The experiment will show the effectiveness of the proposed JSD weighting over the previously proposed SNR mechanism. The last experiment shows the effectiveness of our proposed TDCoW and the importance of context in BU feature weighting for target detection.

6.1 Experiment 1: bottom-up saliency performance

To see the effectiveness of the BU features used along with the global IDM approach of our model for saliency detection, two benchmark saliency datasets were used. The first dataset called ASD or MSR-1000 is the leading benchmark datasets for saliency used by most researchers in this area [2]. It consists of 1000 images, each containing one salient object. The second dataset of 300 images is the SOD dataset, which is a collection of salient object based on Berkeley Segmentation Dataset (BSD) [29]. According to the saliency survey paper [6], this is one of the most difficult and challenging datasets for salient object detection.

The precision-recall curve is used to evaluate the performance at different threshold values. The proposed method that utilizes different features and an IDM-based centre-surround mechanism called IDM (multi-features) or simply IDM (Multi) is compared with 12 state-of-the-art fixation and saliency detection techniques as well as 7 classical techniques. The baseline technique to compare with is the Itti model. We are also comparing the proposed method with one of our previously proposed models for saliency detection called IDM-PCA [33] that utilizes the PCA for dimensionality reduction. The rest of the techniques are abbreviated as follows:

- Classical: MSS [1], CA [39], Rare [35], SWD [10] and GBVS [15].
- State of the art: BSL [40], CAU [32], RRWR [22], UFO [19], SIA(GC) [8], RC and HC [9], QCUT [3], IILP [24], HDCT [21], GMR [44] and FET [14].

Figure 8 shows the precision-recall curves for both the datasets. The curves show that IDM (Multi) outperforms all classical techniques in terms of saliency detection. IDM-PCA has a good performance on both the datasets; however, it uses a single colour feature with dimensionality reduction as opposed to IDM (Multi) which uses various effective and efficient features. In addition, the results also suggest the effectiveness of the features and the proposed IDM-based

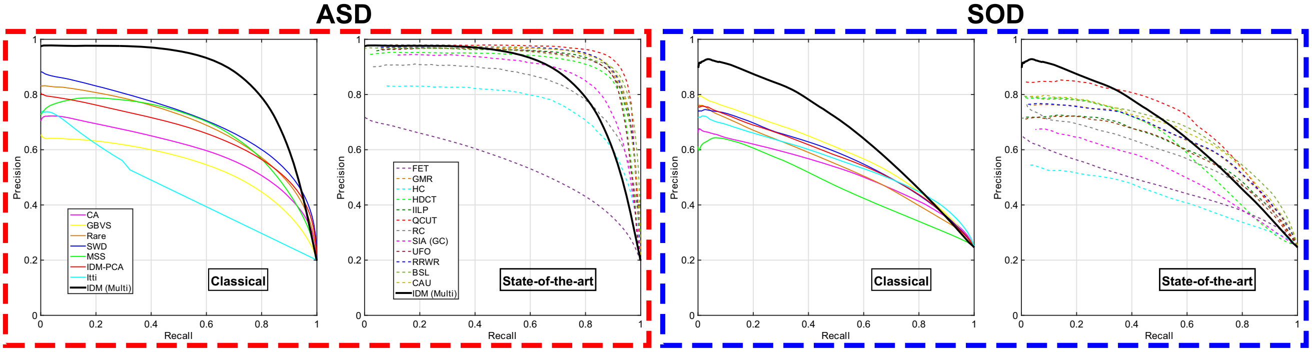


Fig. 8 Precision-recall performance comparison of various classical and state-of-the-art BU techniques with the proposed IDM (Multi-features) model for bottom-up saliency detection on ASD and SOD datasets. (The figure is best viewed in *colour*)

centre-surround mechanism over those being used by Itti's model.

Now comparing our BU model with state-of-the-art techniques, it is obvious that our method has a very high precision value (higher than all other techniques) at low recall values (approximately below 0.4 which corresponds to low threshold) on the ASD dataset. The precision after this point degrades considerably. There are two reasons for this degradation in performance; first it is observed that mostly the saliency maps generated by the proposed method partially highlight the salient object particularly when the size of the salient region is large (a typical characteristic of the salient objects contained in this dataset). Secondly, most of the regions which are highlighted by the proposed technique exhibit low intensity values compared to other techniques. For this reason, when the threshold increases, the precision becomes low as the true positive value is small.

In the more difficult SOD dataset, it can be observed that the proposed technique outperforms all other state-of-the-art techniques considerably for low recall values (approximately 0.5 and below). For high recall values, the degradation in performance is less obvious compared to the degradation occurring on the ASD dataset.

To visualize the maps generated by the proposed BU technique and other state-of-the-art techniques, Fig. 9 shows 8 sample images from ASD and SOD datasets. The saliency maps generated by the all the techniques except for FET, HC and RC exhibit very high precision and low false negatives. Hence on this dataset, and from its sample images we can see that almost all the techniques including the proposed IDM (Multi-features) were able to detect the salient object accurately when compared to the groundtruths (see second row of the figure). On the other hand, from the SOD dataset, we selected four challenging images in which they contain either more than one salient region to be detected (e.g., the last two images in this dataset) or the background and the salient region exhibit similar visual features (e.g., the first and the last image in this dataset).

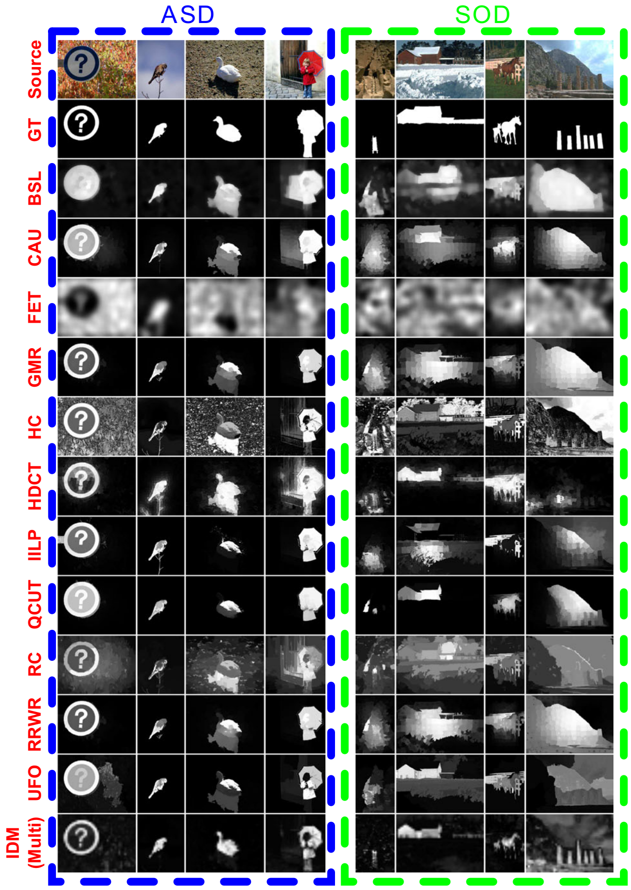


Fig. 9 Qualitative comparison of the bottom-up saliency maps generated by various state-of-the-art techniques and the proposed IDM (Multi-features) on ASD and SOD datasets.

In the first image from this dataset, according to the groundtruth, the salient region is the ladder in front of the rocky mountain. Note that the visual attribute of the background (mountain) and the salient region (ladder) are highly correlated. As a result, the pixel association and propagation-based models such as CAU, GMR, IILP and RRWR assign

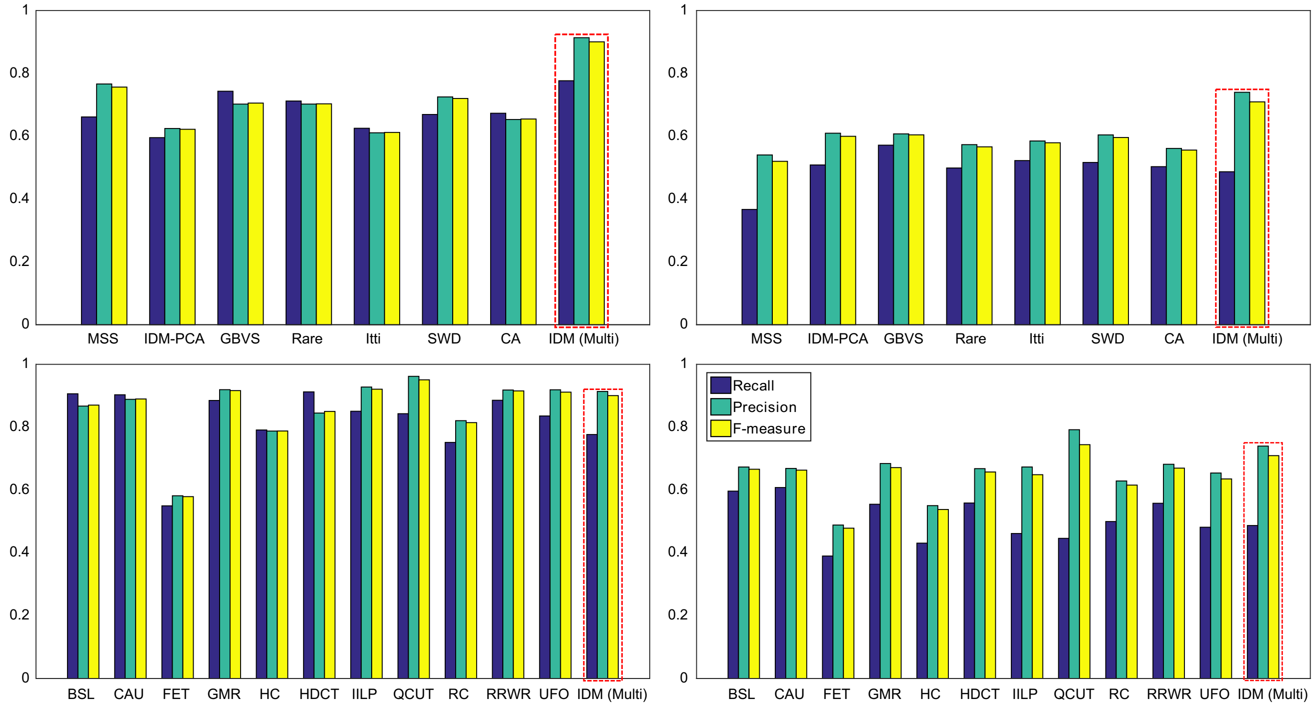


Fig. 10 Segmentation evaluation of various models and the proposed model in terms of precision, recall and F-measure values on ASD (the *first column*) and SOD (the *second column*) datasets. The *first row* represents the comparison with classical techniques and the *second row* for state-of-the-art techniques

same pixels association values to the background and the salient region. Contrast-based methods such as RC and HC also fail to produce satisfactory results as the contrast attributes between the two regions are similar. Even high-level features such as objectness in UFO were not able to separate the two regions. The best effort on this image was delivered by our proposed model (see the last row of the figure). Using an information theoretic approach on various low and mid-level features, the model was able to highlight the salient region which deems to have certain irregularities or element of surprise from these features which are captured by the model. Similar visual analysis is applicable on rest of the images from this dataset.

The proposed BU saliency model although highlights part of the salient object and mostly with low intensity, however, it retains the contour or the over all structure of the object. This is because one of the features it uses is the MSS frequency content based on the method proposed in [1]. As mentioned earlier, MSS technique has a very good segmentation performance on benchmark datasets [6]. Hence, to demonstrate the segmentation capability of our proposed BU model, we applied mean shift adaptive segmentation approach proposed in [2] on the generated saliency maps to produce segmented binary saliency maps. The precision, recall and F-measure are calculated from each saliency map on a particular dataset and averaged over all the images in the dataset. F-measure is another accuracy of detection measure which combines recall and precision metrics and it is given as

$$F_\beta = \left(1 + \beta^2\right) \cdot \frac{\text{precision} \cdot \text{recall}}{\left(\beta^2 \cdot \text{precision}\right) + \text{recall}} \tag{6}$$

where β is an importance factor for weighting either precision or recall. A common value for β is 0.3 [2].

From Fig. 10, we can clearly see that in both datasets, our proposed model has low recall values compared to other classical and state-of-the-art techniques. On the contrary, our model exhibits high precision and in turn F-measure values because precision is weighted more (i.e., assigned more importance) than recall in Eq. (6). In ASD dataset, the precision and F-measure values are comparable to most of the state-of-the-art techniques. In SOD dataset, our model has second largest precision and F-measure values after QCUT technique. Low recall values suggest that there are two many false positive regions within the segmented saliency map. As discussed earlier, this behaviour is obvious as the salient regions detected by the proposed BU model are highlighted with low saliency values. This could be due to simple feature map normalization and integration which is implemented by our model. However, the potential strength of the model lies with its ability to detect the salient object/s with high precision.

Since our BU model has a good performance for BU saliency detection, it is expected to perform well as a platform for TD saliency as the weights are assigned directly to these features.

6.2 Experiment 2: JSD weighting performance

In this experiment, the proposed JSD weighting calculation method is compared with the previously proposed SNR for TD feature weighting. For this experiment, we have used the same setup being followed by the authors in [30] (i.e., same features and centre-surround mechanism based on Itti's model) to avoid any additional processing biasness. The objective of this experiment is to show that JSD is a better option for feature weight calculation than SNR.

The TD weights are calculated with the two approaches once with the proposed JSD method and with the SNR method. For the testing, we have created our own challenging dataset of cricket ball as the target object. The dataset consists of 400 images which are taken indoors and outdoors with variation in size, illumination, distracting objects and background. In addition, some internet images containing the cricket ball are taken to construct the dataset. The reason for including internet images is because we wanted to have some natural images containing target cricket ball (e.g. in well-known cricket grounds, with players, from different matches, etc.). The created dataset is split into two groups of 200 images each, where the first set contains images in which the target object is salient and the second in which the target is non-salient and distracted by other objects. We refer to them as salient and distractor datasets, respectively.

For this experiment, we have randomly selected 100 images (50 from each set) in total from both sets to construct a test to train instance ratio of 1:4 (a common ratio followed in machine learning techniques). The experiment is repeated 10 times to have a random train/test images selection at each run. An average result is obtained in the form of precision-recall curves. It is important to mention here that the weights are calculated for the training images separately in the same way as explained in Sect. 5.2. These weights are averaged to obtain a final single set of feature weights which are in turn universally applied on the testing images as there is no contextual or clustering involved in this experiment. As mentioned earlier, the objective of this experiment is only to show that JSD has is a better choice than SNR for weight calculation.

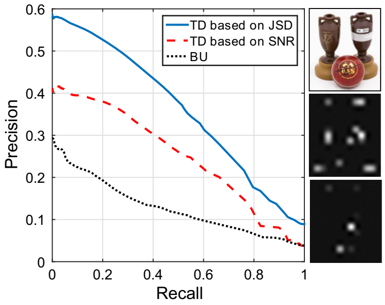


Fig. 11 Comparison between JSD and SNR-based weight calculation methods. The *right column* shows an example of a test image (*top image*) and the saliency map generated when applying SNR (*middle image*) and JSD (*bottom image*) as feature weight calculation procedure

Figure 11 shows the average precision-recall curve for both the JSD and SNR-based TD weighting along with the BU (i.e., no weighting) option. As expected, the highest precision value for the BU approach is very poor (approximately 30 %). This is due to the fact that 50 % of the test images are those in which the target is non-salient, and hence poorly detected by the BU approach. As it is evident, the SNR based TD weighting has improved the detection but only by approximately 10 % of the maximum precision value. On the other hand, the JSD approach has a maximum precision value of 58 % and clearly outperforms the SNR-based approach. Note that the precision is still low due to the absence of contextual information when generating the feature weights. For visual quality, the figure also displays a single image and the obtained TD saliency maps from SNR and JSD, respectively, from top to bottom. As it is evident, the JSD version has fewer false positive regions compared to the SNR one.

From the two experiments mentioned above, we conclude the effectiveness of our proposed features along with the global centre-surround mechanism and the JSD weight calculation procedure for both saliency and target detection. Next with the help of these two results, we demonstrate how feature weighting can improved by incorporating the contextual information.

6.3 Experiment 3: TDCoW for target detection

In this experiment we explore the efficacy of our proposed TDCoW and the importance of the contextual-based clustering. We test our model on four datasets for target object detection. The first two are the salient and distractor datasets discussed earlier for cricket ball target detection. The other two are selected from the Graz-02 dataset which is commonly used for object classification or recognition [31]. The dataset contains images with objects of high complexity and a high intra-class variability on highly cluttered backgrounds. There are three classes in this dataset, however, only two are considered for target detection i.e., bikes and persons as they are more difficult to be classified than the car class.

The images from each of the four datasets were split into equal halves, one for training and the other for testing. In addition, different cluster sizes were used. Figure 12 shows the average area under the curve (AUC) of the receiver operating characteristic (ROC) Analysis achieved when varying the number of clusters in each dataset. It is clear from Fig. 12

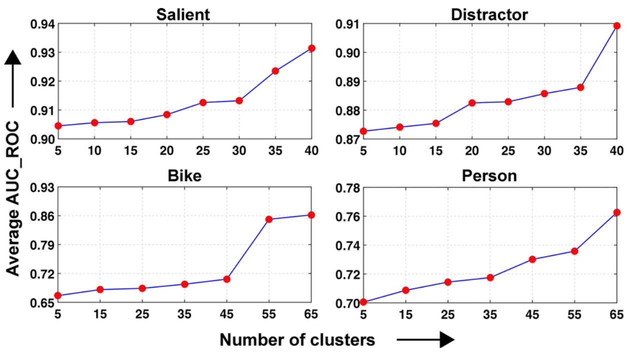


Fig. 12 Cluster size variation affect on the accuracy of the proposed model

that as we increase the number of clusters, we achieve better AUC performance. The drawback of increasing the number of clusters is higher computational load in matching the contextual descriptor of the test image with that of the centroid contextual descriptors of the clusters. To be within a reasonable limit, empirically we have chosen the number of clusters to be 30 for the salient and distractor datasets and 45 for the bike and person datasets.

6.3.1 Quantitative analysis

We test our model by generating the TD saliency maps and finding the accuracy of detection in terms of both precision-recall and F-measure curves.

The comparison is conducted between TDCoW and TD weighting without the contextual information or clustering. Figure 13 shows the obtained results for the salient, distractor, bikes and persons dataset from left to right columns, respectively, where the top row is the precision-recall result and the bottom row is for the F-measure. The proposed TDCoW has a better performance both in terms of precision-recall and F-measure curves in all four datasets. However, some observations and patterns need more elaboration.

Starting with the salient dataset, as we can observe, the BU [i.e., IDM (Multi-features) with flat weighting on the features] has a reasonable performance in the accuracy of detecting the target (see the first column of Fig. 13). This is expected as we have seen the capability of the IDM (Multi-features) in Sect. 6.1 in detecting the salient objects, and this dataset has the target cricket ball being the most salient object in the image. However, when applying the weights without the use of context, the improvement was not that dramatic. This might be due to the fact that averaging the weights over the examples in this dataset yields a nearly uniform distribution of weights. The small improvement may suggest that there is some similarity in structure and context of the images in this dataset. On the other hand, for the proposed TDCoW, we can see a very high improvement in the performance for most of threshold values (see the precision-recall curve in the first row of Fig. 13). This shows the effectiveness of including the context when finding the weights of the features.

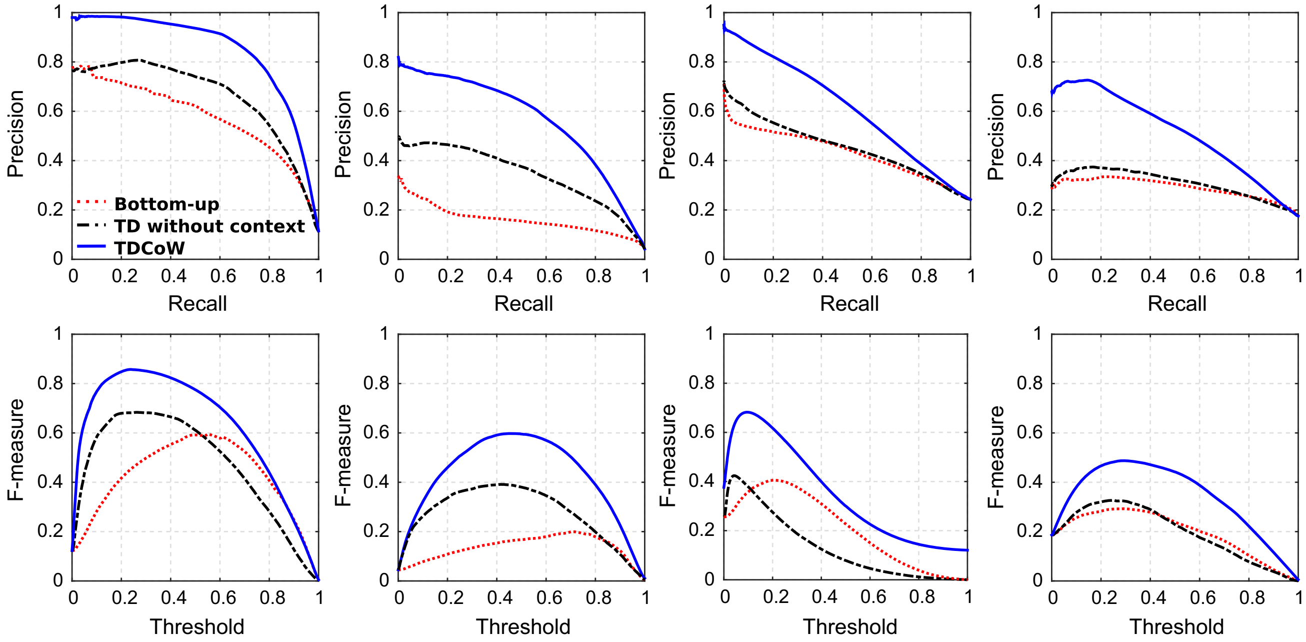


Fig. 13 Precision-Recall and F-measure performance evaluation of TDCoW for the salient, distractor, bikes and persons dataset from *left* to *right*, respectively. The comparison is conducted with the BU model and the TD weighting without context. The *top row* is for Precision-Recall and the *bottom* one is for F-measure

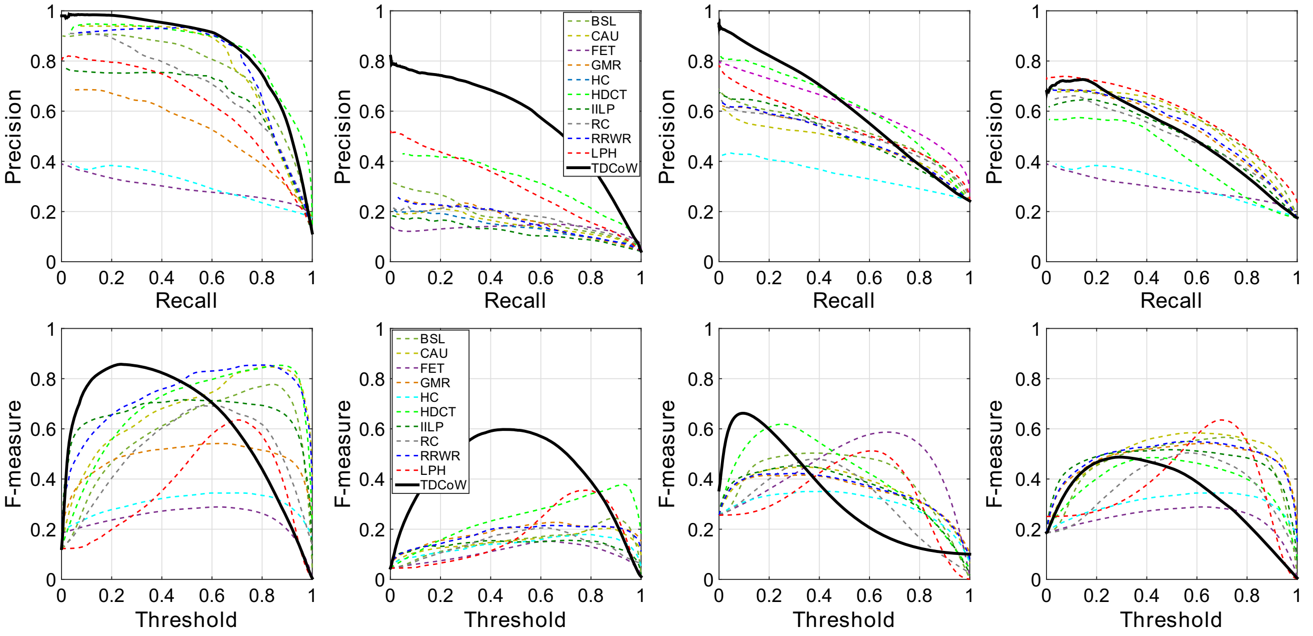


Fig. 14 Precision-recall and F-measure comparison between TDCoW and other state-of-the-art-techniques for the salient, distractor, bikes and persons dataset from *left* to *right*, respectively. The *top row* is for precision-recall and the *bottom* one is for F-measure. (The figure is best viewed in *colour*)

For the distractor dataset (see the second column of Fig. 13), very poor performance both for the BU and TD without context can be observed. This is expected as now the target object is not salient and being distracted by other objects and background variation. Again, the inclusion of context leads to considerable improvement in performance.

The next dataset is the Graz-02 (bike) dataset. Note that there is almost no difference in performance between the BU and the TD without context. In fact, for some high threshold values, the F-measure values are higher for the BU than the TD without context. This suggests that with the absence of context for this dataset in particular, the averaging over the examples is merely a random procedure, and since the images in this dataset have a very high inner-class variability in terms of context, the averaging procedure results in a poor performance due to incorrect weighting of the features. This in turn leads to a degradation in performance, as can be seen strikingly in the F-measure graph in third column of Fig. 13.

The TDCoW on this dataset has a prominent performance improvement on small recall or high threshold values. There is a steep drop in the performance as the threshold values increase. This is because the TD maps generated for this dataset exhibit high variation of intensity values when the target object (in this case a bike) is detected. This might be due to the variation in features of the region of interest containing the bike.

For the last dataset i.e., Graz-02 (persons), again the TDCoW outperforms both BU and TD weighting without context. Due to the difficulty of this dataset, both precision and F-measure are lower than in the other datasets. The TD maps generated by the proposed model have more false positive regions than those observed in the previous three datasets. This gives an indication that the current low-level features being used in TDCoW might not be sufficient to describe such targets. However, the contribution of incorporating context into the weighting mechanism remains prominent in this dataset.

Now to compare the proposed TDCoW model with existing state-of-the-art saliency techniques, we again plot the PR and F-measure curves to evaluate the performance of our proposed model. We also compare our model with the model proposed by Judd et al. (LPH). This model is the closest to ours as it learns weights of various features from eye fixation data through SVM classifier. Since the four datasets used in our experiments have segmentation groundtruth, it is not possible to train the weights over these datasets due to lack of availability of eye fixation information. Instead, from the segmented groundtruth region, we performed a random sampling of points to form an ideal eye fixation data so that the model parameters are learnt from the training examples.

The precision-recall and F-measure curves are replotted in Fig. 14 from Fig. 13 for TDCoW to demonstrate the comparison with other state-of-the-art techniques. As before, the first row of Fig. 14 shows the precision-recall performance, whereas F-measure values are plotted in the second row of the figure for all four datasets. On the first dataset (i.e., salient), it is evident that TDCoW has the best performance than rest of the state-of-the-art techniques. Although in this

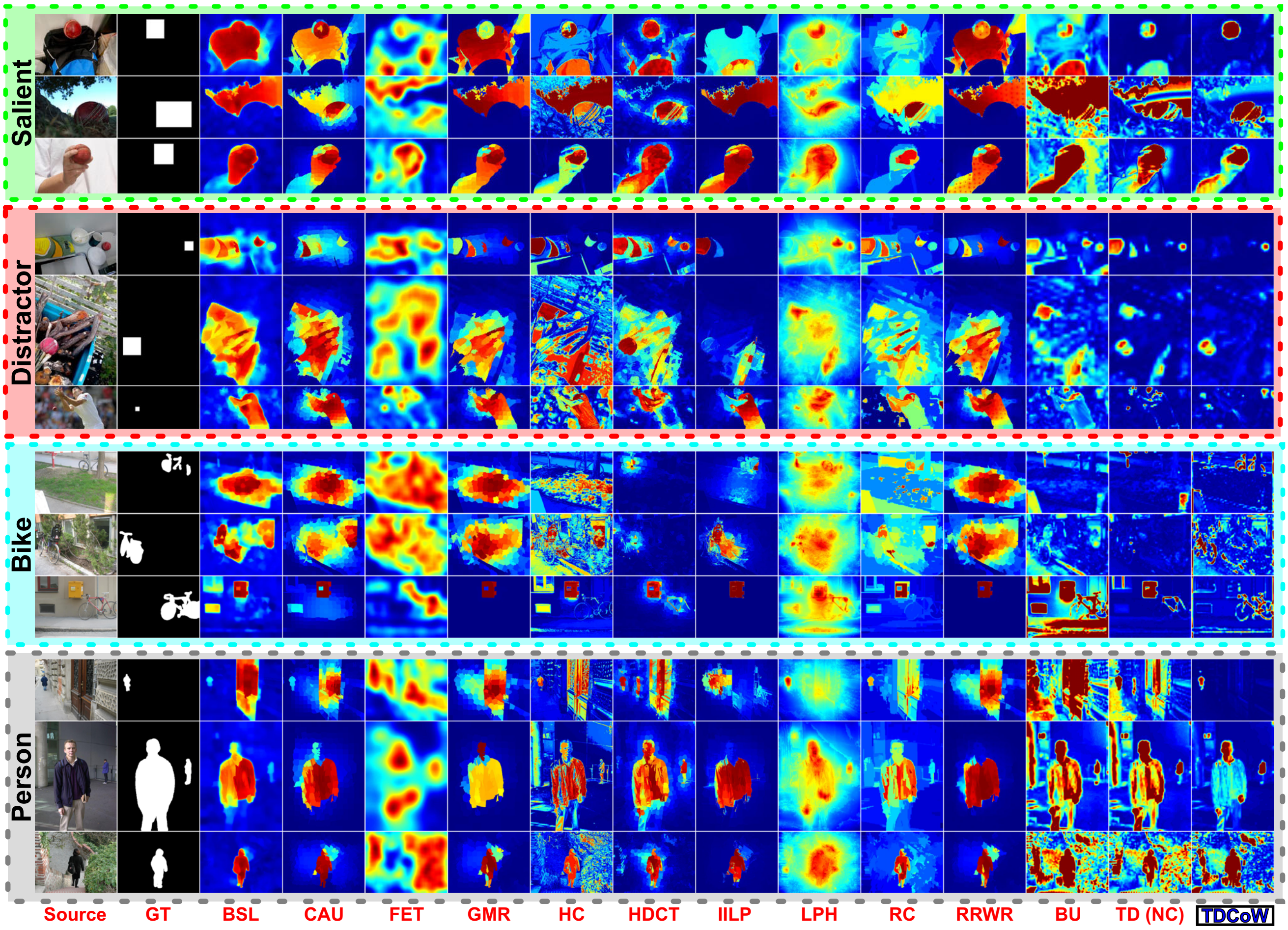


Fig. 15 Qualitative sample images of the proposed TDCoW and other state-of-the-art models. (The figure is best viewed in *colour*)

dataset, the target object to be detected (i.e., the cricket ball) itself is salient, the BU state-of-the-art techniques could not perform as good as our proposed model. The best noticeable performance of TDCoW can be seen for the distractor dataset. A huge performance difference between TDCoW and rest of the state-of-the-art techniques on this dataset confirms the capability of our proposed model in detecting the target object when it is not salient (see the top and bottom rows of column two in Fig. 14). Since the distractor dataset contains distracting object which is mostly salient, the poor performance of these techniques is reflected due to the fact that they falsely detected the most salient region rather than the target object in majority of the examples on this dataset.

In the bike dataset, we can see similar performance by TDCoW to the one acheived on SOD dataset. Most of the images in this dataset contain the target object (i.e., bike) which are salient. As before, TDCoW has the best performance on low threshold values than other techniques but degrades by increasing threshold. Similarly the target object in the person dataset is also salient in most of the images. TDCoW has a moderate performance in this dataset, whereas the best performance is acheived by LPH as the model uses high-level face and pedestrian detectors as features.

6.3.2 Visual analysis

Figure 15 shows representative examples of the saliency maps (shown by heatmaps) generated by various models including our proposed TDCoW model. The last three columns show the maps produced when using BU [i.e., our proposed BU IDM (Multi-features)], TD weighting but without context (TD (NC)), and finally our TD proposed model TDCoW. Three sample images are selected from each dataset. As an example from the salient training image, the ball in the second image although being salient exhibits low contrast and partially occluded by the grass. TDCoW was able to detect the target with high accuracy with a comparable result to HDCT and BSL. Rest of the techniques failed to detect the target precisely. In the distractor sample images, we can see that our model outperforms rest of the techniques in not only detecting the target, but also in producing small false alarm regions.

From the bike dataset, it can be observed that TDCoW was able to detect parts of the bikes and not as a whole object, a point which has been made earlier regarding the detection of the bike dataset. Despite partial detection, the visual results compared to other techniques show very high target detection precision. As an example, in the third image, only our model was able to detect the target with minimal false alarms. On the other hand, rest of the techniques detected the yellow object as it is more salient than the bike. In the person dataset, we can see a reasonable detection performance by the proposed model. For instance, in the last image, the object was detected but with far more false negative regions compared to more accurate results by other techniques.

When comparing our model with the BU version (i.e., no weighting) and the TD weighting without context, we can clearly mark the visual improvement in locating the target object when using context (i.e., by TDCoW model) over the other two approaches (see the last three columns of Fig. 15). In majority of these sample images, we clearly see that pure BU has poor performance in detecting the target, particularly in the distracting and bike dataset. Little improvement is acheived when performing TD weighting of features over all the training examples but without the inclusion of context. Ultimately, upon incorporating the context, the detection performance improvement is obvious. In some situations, for instance (the second and third image of the distraction and persons datasets, respectively), the incorporation of context does not have a significant improvement over the TD weighting without context. In other occasions, a noticeable improvement is acheived when using the context to modify the TD weighting either by increasing the precision in detecting the target (e.g., the first image in the bikes dataset) or by reducing the number of false positive regions (e.g., the first image in the distractor dataset).

6.3.3 Feature weight analysis

Some feature weight statistics are extracted for the proposed model. For the salient and distractor datasets, Fig. 16 shows the ranking of the learned weights. The ranking represents the percentage of examples in a dataset for which a feature weight is positioned at a particular rank. Since there are 17 sub-features in total, the rank is ranged from 1 to 17 indicated by the *x* axis.

As an example, the Principal component FM-1 has the highest rank in terms of weight value (i.e., rank 1) in approximately 30 % of the 100 test images in the salient dataset. This suggests that the PCA FM-1 feature is the most important feature in this dataset for detecting the target.

Now to analyse the weight distribution profile, it can be easily observed that the contrast (particularly red/green and hue channels), both the PCA features and the frequency MSS have the highest weight values particularly in the first four ranking positions in the salient dataset. This result is consistent with some previous work in feature importance for salient object detection. For instance, it has been indicated by many authors the importance of contrast features for BU saliency detection [10,18,39]. Similarly, it has been shown that PCA features play an important role in extracting salient regions [33,45]. The colour, orientation and intensity have no significant contribution as the weights are almost distributed uniformly over these features. This is true as the attention towards a salient object is more concerned with the contrast of the object rather than its colour.

For the centre-bias feature, its importance is not that significant. This might be due to the fact that when the dataset was generated, most of the images that were taken of the target cricket ball did not consider positioning the object in the centre of the image, as usually is the case in most of the saliency datasets.

For the distractor dataset, similar profile of weight ranking can be seen. However, more importance now is on the PCA and contrast features in particular. In addition, colour, edge and orientation have slightly higher weighting than in salient dataset.

6.3.4 Results summary

From the qualitative, quantitative and weight statistics results we conclude that modelling TD saliency by incorporating contextual information plays a major role in performing high-level vision tasks particularly target object detection. Our proposed model TDCoW not only highlights the benefit of using contextual information for target object detection but also demonstrates the superiority of the model over existing state-of-the-art techniques in salient and target object detection.

6.4 Limitations of TDCoW

There are two limitations of the proposed model which need to be explored in future works. The first is whether TD modelling of this kind is sufficient for generic target detection. Although the model utilizes the image context for better feature weight assignment, the knowledge of the target features remains an important factor for more accurate weighting of features.

The second limitation is the processing speed both in the training and testing phases. Generation of a single saliency map for an image size of 400 × 300 pixels requires around 3.5 s. Similarly, weight calculation for all features and CM for a single image takes around 44 s. For 100 training images of an average size of 400 × 300 and with 10 clusters, the CPU time required to perform all the steps in the training phase which includes weight calculation, saliency generation, clustering and contextual descriptor generation is around 4.5 h.

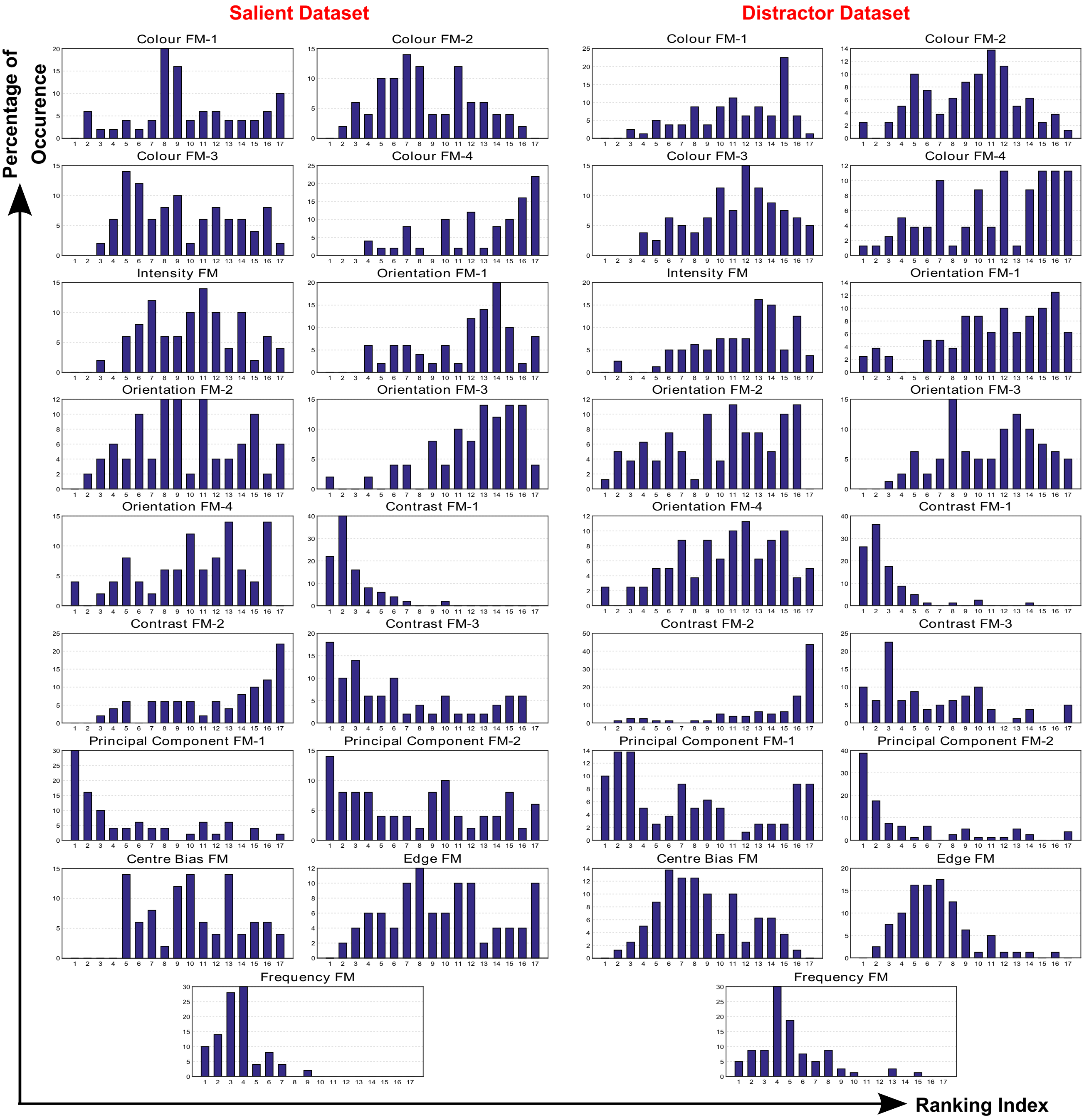


Fig. 16 Feature weight ranking distribution of TDCoW model. The *left half* is the ranking for each feature weight from the salient dataset and the *right* one is for the distractor dataset

In the testing phase, the process involves TD saliency map generation, context generation, context matching and weight selection and takes around 1.5 min to generate the final TD saliency map of an image of size 400 × 300.

Hence, the TDCoW model might not be suitable for real-time target detection applications but could be considered for off-line target detection. Furthermore, to improve the efficiency of the model in the testing phase, the key factor is in the contextual matching. The contextual matching is performed over a large descriptor of distributions which makes the matching process slow. A possible solution could be to reduce the dimensionality of the descriptor for more efficient matching process.

7 Conclusions

Modelling top-down saliency by appropriately weighting the bottom-up features for target detection is a non-trivial research topic in active vision. The major challenge in this research is how to dynamically assign weights to the features. Most of the existing techniques do not consider high-level information within an image when weighting the features. As a result, such learned weights from example images only work when the test images are contextually to the training images.

To overcome this problem, our proposed Top-down Contextual Weighting (TDCoW) model learns contextual structures from the training images and applies them on the test images to dynamically assign weights to the features. Hence, the major contribution of this paper is to highlight the importance of contextual information for top-down saliency modelling by feature weighting for target detection.

The proposed model is tested on four challenging datasets including two self-created datasets of cricket balls and two object classes (bikes and persons) as targets from the Graz-02 dataset. In all datasets, the results show a considerable target detection performance improvement in terms of precision-recall and F-measure values when applying contextual information for feature weighting over feature weighting without context.

References

1. Achanta, R., Susstrunk, S.: Saliency detection using maximum symmetric surround. In: IEEE International Conference on Image Processing 2010 (ICIP), pp. 2653–2656 (2010)
2. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition 2009 (CVPR 09), pp. 1597–1604 (2009)
3. Aytekin, C., Kiranyaz, S., Gabbouj, M.: Automatic object segmentation by quantum cuts. In: International Conference on Pattern Recognition 2014 (ICPR 2014), pp. 112–117 (2014)
4. Benicasa, A.X., Quiles, M.G., Zhao, L., Romero, R.A.F.: Top-down biasing and modulation for object-based visual attention. In: International Conference on Neural Information Processing (ICONIP'13), pp. 325–332 (2013)
5. Borji, A., Itti, L.: Scene classification with a sparse set of salient regions. In: IEEE International Conference on Robotics and Automation (ICRA 2011), pp. 1902–1908 (2011)
6. Borji, A., Sihite, D.N., Itti, L.: Salient object detection: a benchmark. Eur. Conf. Computer Vision **2012**, 414–429 (2012)
7. Borji, A., Cheng, M.M., Jiang, H., Li, J.: Salient object detection: a benchmark. IEEE Trans. Image Process. **24**(12), 5706–5722 (2015)
8. Cheng, M.M., Warrell, J., Lin, W.Y., Zheng, S., Vineet, V., Crook, N.: Efficient salient region detection with soft image abstraction. In: IEEE International Conference on Computer Vision 2013 (ICCV 13), pp. 1529–1536 (2013)
9. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H.S., Hu, S.M.: Global contrast based salient region detection. IEEE Trans. Pattern Anal. Mach. Intell. **37**(3), 569–582 (2015)
10. Duan, L., Wu, C., Miao, J., Qing, L., Fu, Y.: Visual saliency detection by spatially weighted dissimilarity. In: IEEE Conference on Computer Vision and Pattern Recognition 2011 (CVPR 11), pp. 473–480 (2011)
11. Filipe, S., Alexandre, L.A.: From the human visual system to the computational models of visual attention: a survey. Artif. Intell. Rev. (2013)
12. Fornoni, M., Caputo, B.: Indoor scene recognition using task and saliency-driven feature pooling. In: British Machine Vision Conference 2012 (BMVC 2012) (2012)
13. Frintrop, S.: VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search. PhD thesis (2006)
14. Gu, K., Tong, S.J., Zhai, G., Lin, W., Yang, X., Zhang, W.: Visual saliency detection with free energy theory. IEEE Signal Process. Lett. **22**(10), 1552–1555 (2015)
15. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Advances in Neural Information Processing Systems (NIPS), pp. 545–552 (2006)
16. He, S., Han, J., Hu, X., Xu, M., Guo, L., Liu, T.: A biologically inspired computational model for image saliency detection. In: ACM International Conference on Multimedia 2011 (MM 11), pp. 1465–1468 (2011)
17. Hu, Y., Xie, X., Ma, W.Y., Chia, L.T., Rajan, D.: Salient region detection using weighted feature maps based on the human visual attention model. In: Advances in Multimedia Information Processing (PCM 2004), pp. 993–1000 (2004)
18. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998)
19. Jiang, P., Ling, H., Yu, J., Peng, J.: Salient region detection by UFO: uniqueness, focusness and objectness. In: IEEE International Conference on Computer Vision 2013 (ICCV 13), pp. 1976–1983 (2013)
20. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: IEEE International Conference on Computer Vision 2009 (ICCV 09), pp. 2106–2113 (2009)
21. Kim, J., Han, D., Tai, Y.W., Kim, J.: Salient region detection via high-dimensional color transform. In: IEEE Conference on Computer Vision and Pattern Recognition 2014 (CVPR 14), pp. 883–890 (2014)
22. Li, C., Yuan, Y., Cai, W., Xia, Y., Feng, D.: Robust saliency detection via regularized random walks ranking. In: IEEE Conference on Computer Vision and Pattern Recognition 2015 (CVPR 15), pp. 2710–2717 (2015a)
23. Li, G., Shi, J., Luo, H., Tang, M.: A computational model of vision attention for inspection of surface quality in production line. Mach. Vision Appl. **24**(4), 835–844 (2013)
24. Li, H., Lu, H., Lin, Z., Shen, X., Price, B.: Inner and inter label propagation: salient object detection in the wild. IEEE Trans. Image Process. **24**(10), 3176–3186 (2015b)
25. Li, Y., Hou, X., Koch, C., Rehg, J.M., Yuille, A.L.: The secrets of salient object segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition 2006 (CVPR 06), pp. 4321–4328 (2014)
26. McMains, S., Kastner, S.: Interactions of top-down and bottom-up mechanisms in human visual cortex. J. Neurosci. **31**(2), 587–597 (2011)
27. McMains, S.A., Kastner, S.: Visual attention. Encyclopedia of Neuroscience, pp. 4296–4302 (2009)
28. Mitri, S., Frintrop, S., Pervölz, K., Surmann, H., Nüchter, A.: Robust object detection at regions of interest with an application in ball recognition. In: IEEE International Conference on Robotics and Automation 2005 (ICRA 2005), pp. 125–130 (2005)
29. Movahedi, V., Elder, J.H.: Design and perceptual validation of performance measures for salient object segmentation. In: IEEE

Computer Society Workshop on Perceptual Organization in Computer Vision (POCV), pp. 49–56 (2010)
30. Navalpakkam, V., Itti, L.: An integrated model of top-down and bottom-up attention for optimizing detection speed. In: IEEE Conference on Computer Vision and Pattern Recognition 2006 (CVPR 06), vol. 2, pp. 2049–2056 (2006)
31. Opelt, A., Pinz, A., Fussenegger, M., Auer, P.: Generic object recognition with boosting. IEEE Trans. Pattern Anal. Mach. Intell. **28**(3), 416–431 (2006)
32. Qin, Y., Lu, H., Xu, Y., Wang, H.: Saliency detection via cellular automata. In: IEEE Conference on Computer Vision and Pattern Recognition 2015 (CVPR 15), pp. 110–119 (2015)
33. Rahman, I.M.H., Hollitt, C., Zhang, M.: Information divergence based saliency detection with a global center-surround mechanism. In: International Conference on Pattern Recognition 2014 (ICPR 14), pp. 3428–3433 (2014)
34. Rasolzadeh, B., Targhi, A.T., Eklundh, J.O.: An attentional system combining top-down and bottom-up influences. In: International Workshop on Attention in Cognitive Systems (WAPCV 2007), pp. 123–140 (2007)
35. Riche, N., Mancas, M., Duvinage, M., Mibulumukin, M., Gosselin, B., Dutoit, T.: RARE2012: a multi-scale rarity-based saliency detection with its com parative statistical analysis. Signal Process. Image Commun. **28**(6), 3114–3124 (2013)
36. Rothkopf, C.A., Ballard, D.H., Hayhoe, M.M.: Task and context determine where you look. J. Vision **7**(14), 1–20 (2007)
37. Siagian, C., Itti, L.: Rapid biologically-inspired scene classification using features shared with visual attention. IEEE Trans. Pattern Anal. Mach. Intell. **29**(2), 300–312 (2007)
38. Spotorno, S., Malcolm, G.L., Tatler, B.W.: How context information and target information guide the eyes from the first epoch of search in real world scenes. J. Vision **14**(2 (Article 7)), 1–21 (2014)
39. Stas, G., Lihi, Z.M., Ayellet, T.: Context-aware saliency detection. IEEE Trans. Pattern Anal. Mach. Intell. **34**(10), 1915–1926 (2012)
40. Tong, N., Lu, H., Ruan, X., Yang, M.H.: Salient Object Detection via Bootstrap Learning, pp. 1884–1892 (2015)
41. Torralba, A., Olivia, A., Castelhano, M.S., Henderson, J.M.: Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychol. Rev. **113**(4), 766–786 (2006)
42. Treisman, A., Gelade, G.: A feature integration theory of attention. Cogn. Psychol. **12**, 97–136 (1980)
43. Uijlings, J.R.R., van de Sande, K.E.A., Gevers, T., Smeulders, A.W.M.: Selective search for object recognition. Int. J. Computer Vision **104**(2), 154–171 (2013)
44. Yang, C., Zhang, L., Lu, H., Yang, M.: Saliency detection via graph-based manifold ranking. In: IEEE Conference on Computer Vision and Pattern Recognition 2013 (CVPR 13), pp. 3166–3173 (2013)
45. Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.W.: SUN: a bayesian framework for saliency using natural statistics. J. Vision **8**(7 (Article 32)), 1–20 (2008)
46. Zhaoping, L., Frith, U.: A clash of bottom-up and top-down processes in visual search: the reversed letter effect revisited. J. Exp. Psychol. Hum. Percept. Perform. **37**(4), 997–1006 (2011)
47. Zhaoping, L., Guyader, N.: Interference with bottom-up feature detection by higherlevel object recognition. Curr. Biol. **17**, 26–31 (2007)
48. Zhou, L., Zhou, Z., Hu, D.: Scene classification using a multi-resolution bag-of-features model. Pattern Recognit. **46**(1), 424–433 (2013)

Ibrahim Rahman received a BE degree in computer systems from NED University of Engineering and Technology, Karachi, in 2002 and a Master of Science degree with specialization in computer networks and communication in 2007 from Sir Syed University of Engineering and Technology (SSUET), Karachi. He is a permanent faculty member in the department of computer engineering at SSUET. Currently he is pursuing his PhD in computer vision and image processing from the Victoria University of Wellington. He is working on developing efficient and effective techniques for object recognition using various active vision and machine learning models. His research interest includes machine learning, active vision, human perception and signal processing.

Christopher Hollitt completed a BE(Hon) in Electrical and Electronic Engineering and a BSc(Hon) in Physics and Theoretical Physics at the University of Adelaide in 1994 and 1996, respectively. In 2007 he received a PhD at the same institution, having studied the control of high precision optical systems for gravitational wave detection. His current research work is in machine perception, image processing and robot control. The work spans a wide area, from fundamental problems in feature recognition, through techniques for efficiently utilizing the limited sensory resources of a robot system, to high-level applications of image processing. Dr Hollitt is a senior lecturer in the School of Engineering and Computer Science at the Victoria University of Wellington.

Mengjie Zhang (M'04-SM'10) received the B.E. and M.E. degrees from Artificial Intelligence Research Center, Agricultural University of Hebei, Hebei, China, and the Ph.D. degree in computer science from RMIT University, Melbourne, VIC, Australia, in 1989, 1992, and 2000, respectively. Since 2000, he has been with the Victoria University of Wellington, Wellington, New Zealand, where he is currently Professor of Computer Science, Head of the Evolutionary Computation Research Group, and the Associate Dean (Research and Innovation) in the Faculty of Engineering. His current research interests include evolutionary computation, particularly genetic programming, particle swarm optimization, and learning classifier systems with application areas of image analysis, multiobjective optimization, classification with unbalanced data, feature selection and reduction, and job shop scheduling. He has published over 400 academic papers in refereed international journals and conferences. Prof. Zhang has been serving as an Associated Editor or Editorial Board Member for five international journals (including IEEE Transactions on Evolutionary Computation, Evolutionary Computation Journal) and as a Reviewer of over 20 international journals. He has been serving as a Steering Committee Member and a Program Committee Member for over 80 international conferences. He has supervised over 50 postgraduate research students. He is the Chair of the IEEE CIS Evolutionary Computation Technical Committee, a member of the IEEE CIS Intelligent Systems and Applications Technical Committee, a Vice-Chair of the IEEE CIS Task Force on Evolutionary Computer Vision and Image Processing, a Vice-Chair of the IEEE CIS Task Force on Evolutionary Computation for Feature Selection and Construction, a member of IEEE CIS Task Force of Hyper-heuristics, and the Founding Chair for IEEE Computational Intelligence Chapter in New Zealand.