

HARF: Hierarchy-associated Rich Features for Salient Object Detection

Wenbin Zou

Shenzhen Key Lab of Advanced Telecommunication and Information Processing
College of Information Engineering, Shenzhen University

zouszu@sina.com

Nikos Komodakis

Universite Paris-Est, Ecole des Ponts ParisTech

nikos.komodakis@enpc.fr

Abstract

The state-of-the-art salient object detection models are able to perform well for relatively simple scenes, yet for more complex ones, they still have difficulties in highlighting salient objects completely from background, largely due to the lack of sufficiently robust features for saliency prediction. To address such an issue, this paper proposes a novel hierarchy-associated feature construction framework for salient object detection, which is based on integrating elementary features from multi-level regions in a hierarchy. Furthermore, multi-layered deep learning features are introduced and incorporated as elementary features into this framework through a compact integration scheme. This leads to a rich feature representation, which is able to represent the context of the whole object/background and is much more discriminative as well as robust for salient object detection. Extensive experiments on the most widely used and challenging benchmark datasets demonstrate that the proposed approach substantially outperforms the state-of-the-art on salient object detection.

1. Introduction

Saliency detection is important to many applications, such as object segmentation, object recognition, content-based image retrieval, and adaptive image/video coding. The original task of saliency detection aims to predict fixation points in an image, where the earliest study on this topic is motivated by the observation from cognitive scientists that the inherent visual attention mechanism enables humans to identify rapidly and effortlessly the visually outstanding (salient) regions/objects in complex scenes.

Following the fixation prediction, saliency detection has

been recently extended to mean highlighting the whole salient object in an image. The focus of the present paper is exactly on this task (usually also referred to as salient object detection in the computer vision literature), which is essentially equivalent to solving a binary foreground/background segmentation problem.

A common characteristic of most existing methods for salient object detection is that they generally simplify the original images by partitioning them into blocks or segmenting them into regions through image segmentation algorithms or pixel clustering methods. This is done for efficiency but, most importantly, for computing visual features of wider image support with the hope that these features will thus be more robust. The fixed block partition inevitably merges object pixels into background for those blocks surrounding the object boundaries, while image segmentation methods are typically boundary-preserved and ensure that the pixels within the same region share certain visual characteristics. Thus the state-of-the-art salient object detection models typically compute visual features based on regions for saliency evaluation, where it is important to note that the robustness and richness of these features also determine to a large extent performance of saliency detection.

However, one problem from this is that the features extracted from small regions might be not sufficiently discriminative for detecting salient objects in complex scenes. A suggestion might be to adjust the segmentation parameters so that an object is composed of very few regions to facilitate the salient object detection task. Unfortunately, natural images may contain a variety of complex scenes, and the state-of-the-art image segmentation methods are still far from separating the whole objects with the well-defined boundaries from background regions. For some images, adjusting the segmentation parameters to decrease the number of regions may result in under-segmentation, where salient object regions are merged into background, and lead to in-

This work was supported by the NSFC projects (No. 61401287 and No. 61472257), the EC project FP7-ICT-611145 ROBOSPECT and the Shenzhen Key Project for Foundation Research (No. JC201105170613A).

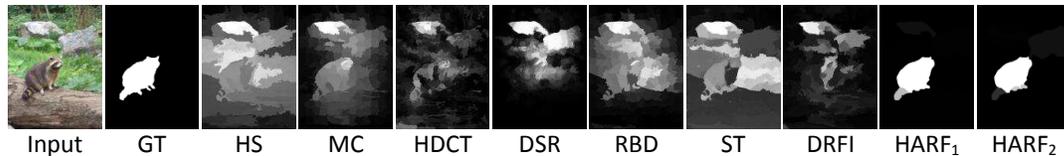


Figure 1. Saliency maps generated by seven state-of-the-art saliency models, and by the proposed model with HARF₁ and HARF₂ description respectively. Most previous models suffer from their limited robustness to highlight the whole salient object in complex scenes.

accurate salient object detection. As a result, even the state-of-the-art saliency models still have significant difficulties in completely highlighting salient objects in complex scenes (see Figure 1). Therefore, given the above discussion, a fundamental question that needs to be addressed is the following: “how can we extract features that are both more *robust* and also *richer* based on the over-segmented regions for salient object detection?”

To that end, this paper proposes a novel hierarchy-associated feature construction framework for salient object detection, which integrates various elementary features from both the target region and super-regions in a hierarchy. Our hypothesis is that features computed in this manner are able to represent the context of the entire object/background and are much more discriminative as well as robust for salient object detection. In such a context, the paper also introduces the use of enriched elementary features, which are computed from the outputs of multiple hidden layers of a deep convolutional neural network (CNN). By regional contrasts evaluation, these CNN features and other typical elementary features are effectively incorporated into the proposed feature construction framework to generate *hierarchy-associated rich features* (HARF). With such a rich feature representation, we are able to cast saliency detection as a regression problem for which a boosted predictor is trained to estimate regional saliency scores. Extensive experiments on the most widely used benchmark datasets demonstrate that the HARF representation achieves higher performance of salient object detection and the proposed approach substantially outperforms the state-of-the-art models.

In summary, the paper contributions are as follows:

1. We propose a novel hierarchy-associated feature construction framework for salient object detection that allows much more robust and discriminative features, which utilize information from multiple image regions.
2. Furthermore, we introduce the use of multi-layered deep learning features, which are incorporated into the above framework through a compact feature integration scheme that allows to efficiently use multiple elementary features of high-dimensionality.
3. Last, we show that the proposed approach outperforms the state-of-the-art saliency models by a large margin, both quantitatively and qualitatively.

2. Related work

Regarding the early problem of fixation prediction, Itti et al. [13] proposed a well-known saliency model which was implemented based on the biological attention mechanisms and feature integration theories. In this model, elementary features, e.g., color and luminance computed from different scales, are integrated using a center-surround operator to generate the saliency map, in which visually salient points are highlighted, as the prediction of fixations. After that, a number of fixation prediction models are proposed (e.g., [12, 25]). A comprehensive survey on the fixation prediction models can be found in [5].

Concerning the salient object detection problem, which is the focus of this paper, in [36] it was defined as a binary segmentation problem for application to object recognition. Since then, plenty of saliency models have been proposed for detecting salient objects in images based on various theories and principles, such as information theory [37], graph theory [15, 23, 41], statistical modeling [8, 34, 39], low-rank matrix recovery [35, 44], partial differential equations [26], and machine learning [17, 20, 27, 29, 42]. Moreover, a variety of effective measures and priors are explored to achieve a higher performance of salient object detection, e.g., local and global contrast measures [1, 9, 16, 28, 30, 33, 40], center prior [19], boundary connectivity prior [38, 43, 44], focusness prior [18, 22], objectness prior [7, 14] and background prior [15, 24, 26, 41]. Apart from detecting salient objects in a single image, salient object detection also has been extended to identifying common salient objects shared in multiple images and video sequences. For a comprehensive survey on salient object detection models, the interested reader is referred to [3].

Some recent models have exploited hierarchical architectures for salient object detection. In [41], hierarchical inference framework is proposed to fuse multi-scale saliency cues, while in [28] the most confident regions in a saliency tree are selected to improve the performance of salient object detection. In contrast, we focus on extracting discriminative and robust features through hierarchical representation for detecting salient objects in complex scenes.

3. Hierarchy-associated feature construction framework

This section describes how the hierarchy-associated features are constructed. This construction starts by first build-

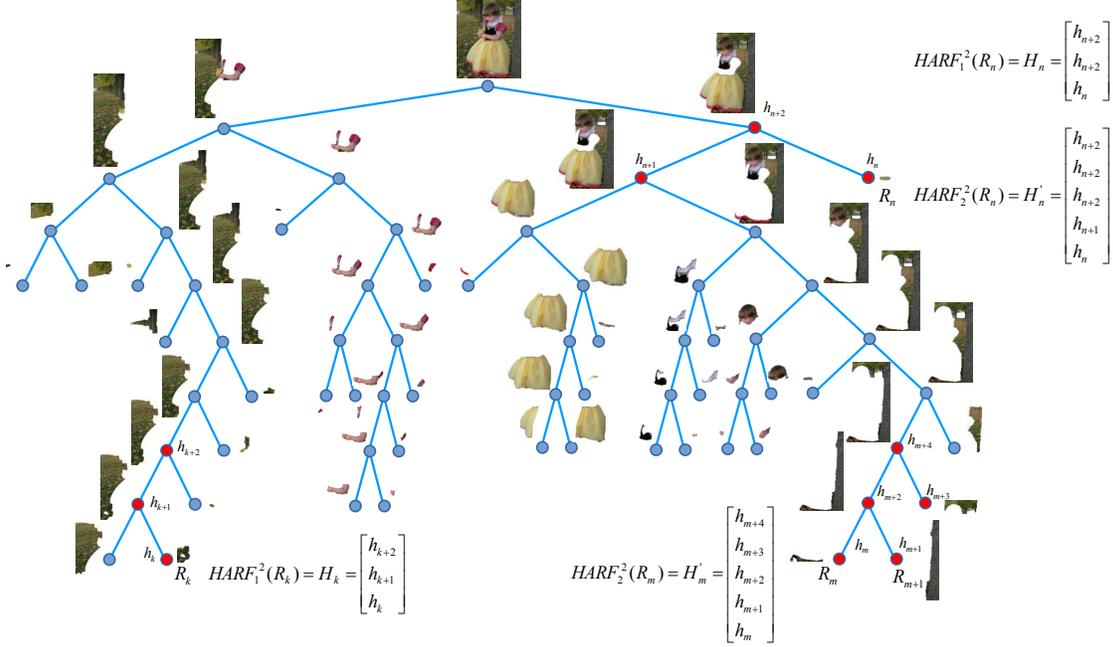


Figure 2. Hierarchy-associated feature construction: this figure illustrates how to compose basic regional features for constructing HARF descriptors for regions R_k , R_m and R_n (here we use h_i to denote the basic regional feature of region R_i).

ing a binary segmentation tree (Section 3.1), which allows us to extract hierarchical image regions and to systematically analyze their neighborhood relationships. Section 3.2 explains how HARF features are generated by combining multiple basic regional features, whereas Section 3.3 presents the computation of the basic regional features that is achieved through a compact integration of multiple elementary features.

3.1. Binary segmentation tree generation

To construct a binary segmentation tree for an input image, we exploit the gPb (globalized probability of boundary based contour detection) method [2] to generate an ultrametric contour map (UCM), which contains a set of real-valued contours. We normalize the UCM into the range of $[0, 1]$ and generate a segmentation containing approximately 100 regions by thresholding with appropriate contour values. Then the generated regions $R_i (i = 1, \dots, K)$ are the basis to construct a binary segmentation tree, in which each leaf node represents a region and each non-leaf node represents a super-region.

For generating super-regions in the segmentation tree, the merging order is determined by contour values of adjacent regions, *i.e.*,

$$R_s = \arg \min C(R_n, R_j), R_n \in \mathcal{N}_j \quad (1)$$

where \mathcal{N}_j denotes the set of neighbors of the target region R_j , and $C(R_n, R_j)$ denotes the contour value between regions R_n and R_j in the UCM. This means that the pair of

regions with the lowest contour value are merged first to generate a new super-region R_s . The merging process is performed iteratively to generate the binary segmentation tree until the final two super-regions are merged to compose the entire image. Figure 2 illustrates the segmentation tree generated for the example image.

3.2. Hierarchical feature construction

Based on the constructed binary segmentation tree illustrated in Figure 2, we propose a novel feature construction framework from the following observations.

To begin with, the image regions (segments) at leaf-nodes in the segmentation tree are typically small, even humans are not able to efficiently recognize most of them while only looking at a single one. Thus the extracted features only from the small regions would be not sufficiently discriminative for salient object detection. Second, the whole object (the little girl) is not completely separated from the background in any regions or super-regions in the segmentation tree. However, the main parts of the object become increasingly clear when we look from the leaf-nodes to the root through the segmentation hierarchy, such as the arms, the skirt and the head. This suggests that the binary hierarchical representation is able to capture global context of objects. Based on the above observations, we propose the following two methods to compute hierarchy-associated rich features (HARF) for regions at leaf-nodes.

HARF₁: In this case, we propose combining basic regional features, whose exact form will be detailed in the

next section, from both the local region and more global super-regions to form HARF. Specifically, basic regional features are computed first for both the target region to be represented and for super-regions corresponding to its β ancestor nodes in the segmentation tree. Then the extracted basic regional features from them are stacked into a single feature vector. For abbreviation, the feature extracted in this form is denoted as $\text{HARF}_1^\beta(R)$, where R represents the target region and β is the number of levels of super-regions in the segmentation hierarchy included to compute HARF. Therefore, assuming that the basic regional feature for each region or super-region is represented by a d -dimensional feature vector, the dimensionality of the generated HARF_1^β feature is $d + d \times \beta$. An illustration of computing feature $\text{HARF}_1^\beta(R_k)$ is shown in Figure 2.

HARF₂: Alternatively, given a region R_m , we propose to compute a richer HARF feature, denoted by $\text{HARF}_2^\beta(R_m)$, by considering not only ancestor super-regions but also sibling regions (in the binary segmentation tree). More specifically, if $\{R_m^{(i)}\}_{i=1}^\beta$ are the β ancestor regions of R_m , to compute $\text{HARF}_2^\beta(R_m)$ we then consider all of the regions used in $\text{HARF}_1^\beta(R_m)$ plus regions $\text{sibling}(R_m)$ and $\{\text{sibling}(R_m^{(i)})\}_{i=1}^{\beta-1}$, where $\text{sibling}(R)$ denotes the sibling region of R . The basic regional features from all these regions are stacked to form $\text{HARF}_2^\beta(R_m)$, which has dimensionality $d + 2d \times \beta$. Figure 2 provides a visual illustration of constructing feature $\text{HARF}_2^\beta(R_m)$.

Notice that, the whole image at the root-node is not included for HARF computation because it certainly contains both salient object and background regions. Furthermore, not all regions at leaf-nodes have the same levels of hierarchy to reach the root-node, and therefore some regions may not have sufficient levels to compute HARF. In this case, the basic regional feature from the super-region at the highest level is duplicated and combined with other features from lower levels to generate the HARF feature of the target region, such as $\text{HARF}_1^2(R_n)$ and $\text{HARF}_2^2(R_n)$ in Figure 2. Experimentally, 8 levels of hierarchy (*i.e.*, $\beta = 8$) were found to be sufficient for all the images we tested.

Due to the way they have been constructed, the proposed features have the following nice properties:

1. *Context-aware*: HARF is computed from both the target local region and super-regions in a hierarchical segmentation tree, thus global context is encoded for the target region representation.
2. *Discriminative*: Due to its wider support, HARF is more discriminative and robust compared to the traditional features that are computed directly from the target region only.
3. *Efficient*: HARF is a combination of multiple basic regional features, thus its computation is efficient.

3.3. Compact integration of rich elementary features for HARF construction

It remains to describe how the basic regional features are formed (which are used for the HARF construction presented in the previous section). To that end, given a region R_j , its local regional feature \mathbf{h}_j is estimated in two stages: (i) the *elementary features computation*, where (possibly high-dimensional) elementary features of several different types, say $\{\mathbf{f}_j^1, \mathbf{f}_j^2, \dots, \mathbf{f}_j^Q\}$, are computed for region R_j , and (ii) the *compact feature integration*, where the estimated elementary features are compactly integrated to produce two low-dimensional feature vectors, the local regional contrasts $\mathbf{h}_j^l = [h_{j,1}^l; \dots; h_{j,Q}^l]$ and the border regional contrasts $\mathbf{h}_j^b = [h_{j,1}^b; \dots; h_{j,Q}^b]$. These, together with a regional property descriptor \mathbf{h}_j^p (describing geometric properties of region R_j), are used to form the basic local regional descriptor \mathbf{h}_j , *i.e.*, $\mathbf{h}_j = [\mathbf{h}_j^l; \mathbf{h}_j^b; \mathbf{h}_j^p]$.

3.3.1 Using CNN features as rich elementary features

In our framework, different types of elementary features can be used to capture different visual characteristics of an image region. Salient object detection aims to highlight objects that stand out from background regions, which essentially is a contrast modeling problem. Thus most of the previous saliency models define regional saliency based on the regional contrasts of traditional low-level features, such as color and texture. Although such low-level features may be able to address relatively simple scenes, however for more complex scenes they suffer from their limited robustness. Therefore, here we propose to also combine deep learning features from convolutional neural networks (CNN).

CNN-based deep learning features have been recently applied to semantic-level vision tasks, *e.g.*, image classification, object recognition and semantic segmentation, yet they are rarely exploited in saliency detection domain. The CNN generally consists of several convolutional layers and fully-connected layers. Typically, the outputs of the last layer of CNN are used as features in semantic-level tasks, which makes sense for these applications because the last layer carries the highest semantic information. However, salient object detection focuses on highlighting salient objects from background, and does not need to assign semantic labels to any objects. The semantic-sensitive last layer may be insufficient for the low-semantic task. Therefore, we propose to exploit all earlier hidden layers for saliency evaluation.

Specifically, we use the CNN model of [11], which is implemented based the architecture of [21] and contains five convolutional layers and two fully connected layers. The CNN parameters are trained from a large dataset of the 2012 ImageNet large scale visual recognition challenge with image-level annotations. The CNN model requires in-

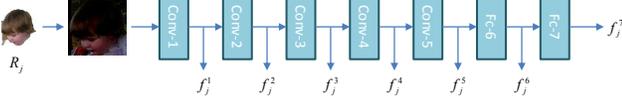


Figure 3. The outputs of each layer of CNN are used as features.

puts of a fixed 224×224 pixel size. Therefore, for regional feature extraction, we first warp the bounding box surrounding the target region to an image of size 224×224 and subtract from it the mean of training images (for normalization), then propagate it through the network. As illustrated in Figure 3, each CNN hidden layer output vector is used as a separate elementary feature, thus leading to 7 different types of elementary features, $\{\mathbf{f}_j^1, \dots, \mathbf{f}_j^7\}$. These are combined with traditional low-level elementary features as in [17], which consist of LBP texture descriptors and color descriptors (mean of color values and color histograms from RGB, HSV and Lab spaces).

3.3.2 Compact feature integration

Let $\{\mathbf{f}_j^1, \dots, \mathbf{f}_j^Q\}$ denote the set of elementary features computed for the target region R_j . Due to their high dimensionality, it is impractical to directly combine all of them into a feature vector. Therefore, for a compact representation and reasonable saliency evaluation, we compute local regional contrasts and border regional contrasts for each elementary feature type as follows.

The local regional contrasts evaluate dissimilarities between the target region R_j and its adjacent neighbors \mathcal{N}_j in terms of each elementary feature (including CNN features, color features and texture features). The local contrast of feature \mathbf{f}_j^q ($q = 1, \dots, Q$) for region R_j is defined as

$$h_{j,q}^l = \sum_{n \in \mathcal{N}_j} \mu_n \lambda_n \mathcal{D}(\mathbf{f}_n^q, \mathbf{f}_j^q) \quad (2)$$

where \mathbf{f}_n^q denotes the feature of neighbor region R_n ; $\mathcal{D}(\cdot)$ is a distance function; μ_n and λ_n are the normalized area weight and spatial weight for neighbor region R_n , respectively. The area weight μ_n is defined based on the assumption that the larger neighboring regions have more contributions to compute the local contrast value $h_{j,q}^l$, *i.e.*,

$$\mu_n = \frac{|R_n|}{\sum_{i \in \mathcal{N}_j} |R_i|} \quad (3)$$

Similarly, those neighboring regions more near to the target region are assigned larger spatial weight λ_n , *i.e.*,

$$\lambda_n = \frac{1/\|\mathbf{x}_n - \mathbf{x}_j\|_2}{\sum_{i \in \mathcal{N}_j} 1/\|\mathbf{x}_i - \mathbf{x}_j\|_2} \quad (4)$$

where \mathbf{x}_j denote the centroid coordinates of target region R_j ; \mathbf{x}_n and \mathbf{x}_i represent the centroid coordinates of neighboring regions R_n and R_i , respectively. The distance function

$\mathcal{D}(\cdot)$ in Eq. (2) is defined according to the used elementary features. For example, the chi-squared distance is typically a good choice for histogram-based features, while for most other features, Euclidean distance can be used.

The border regional contrasts are defined based on the observation in previous work [44] that background regions generally connect with image borders while most object regions do not connect with them. We compute the border regional contrasts between target region R_j and bordering regions R_b ($b = 1, \dots, B$) for each feature \mathbf{f}_j^q . The border regional contrasts may be computed by taking all the bordering regions as a whole or considering them independently. In the former case, the bordering regions are grouped into a single region R_g . Thus the contrast $h_{j,q}^b$ of the elementary feature \mathbf{f}_j^q from region R_j to bordering region R_g is defined as

$$h_{j,q}^b = \mathcal{D}(\mathbf{f}_j^q, \mathbf{f}_g^q) \quad (5)$$

where \mathbf{f}_g^q denotes the feature computed from region R_g . In the latter case, the border regional contrast $h_{j,q}^b$ is computed as the distance between the feature \mathbf{f}_j^q and the feature average of all bordering regions, *i.e.*,

$$h_{j,q}^b = \mathcal{D}(\mathbf{f}_j^q, \frac{1}{B} \sum_{b=1}^B \mathbf{f}_b^q) \quad (6)$$

where \mathbf{f}_b^q ($b = 1, \dots, B$) denotes the features of bordering regions. In our experiments, Eq. (5) is applied to border regional contrasts for color and texture features, while Eq. (6) is used for CNN features. This is because the input to the CNN model needs to be a rectangular image block, which for the case of the grouped border region R_g would lead to using an image block with a big empty region at its interior.

Last, as in [17], a regional property descriptor \mathbf{h}_j^p , which encodes geometric properties and the variances of the elementary features, is also included for region description.

To summarize, the basic regional feature \mathbf{h}_j of region R_j is a combination of local regional contrasts, $\mathbf{h}_j^l = [h_{j,1}^l; \dots; h_{j,Q}^l]$, border regional contrasts, $\mathbf{h}_j^b = [h_{j,1}^b; \dots; h_{j,Q}^b]$, and regional property descriptors \mathbf{h}_j^p .

4. Saliency prediction through regression

With the HARF representation for each leaf region of the binary segmentation tree, we cast salient object detection as a regression problem that predicts the saliency of a region. For the regression, we used the AdaBoost algorithm in our experiments due to its efficiency in both training and testing. Such a method iteratively assembles weak decision trees to generate a single composite strong learner. In this case, given a set of training regions represented by HARF $\{\mathbf{H}_1, \dots, \mathbf{H}_T\}$ along with their ground-truth labeling, we learn a boosted regressor of the form

$$w(\mathbf{H}_k) = \sum_u \alpha_u \mathcal{T}_u(\mathbf{H}_k) \quad (7)$$

where α_u and $\mathcal{T}_u(\mathbf{H}_k)$ are the trained coefficient and the weak decision tree, respectively. We use Eq.(7) to obtain the predicted score of the boosted regressor for a test region R_k , represented by HARF descriptor \mathbf{H}_k . For reasonable saliency precision, we compute the saliency score s_k of R_k by further fitting the output of the boosted regressor into the range of $[0, 1]$ with a sigmoid function, *i.e.*,

$$s_k = \frac{1}{1 + e^{-w(\mathbf{H}_k)}} \quad (8)$$

5. Experiments

We performed an extensive evaluation of our method by applying it to the most commonly used benchmark datasets and by also comparing it to several state-of-the-art salient object detection methods.

5.1. Datasets

We perform experiments on MSRA-B [27], PASCAL-1500 [44] and SOD [32] datasets.

MSRA-B includes 5000 images, most of which contain a single salient object from a variety of scenes. the manually segmented ground truths provided by [17] are used for an accurate evaluation. It should be noted that, the widely used and relatively simpler ASD (or MSRA-1000) dataset [1] is a subset of MSRA-B dataset. Thus ASD dataset is not included for performance evaluation.

PASCAL-1500 includes 1500 images from PASCAL VOC 2012 segmentation challenging [10]. Many images in PASCAL-1500 contain multiple salient objects with various scales, locations and or highly clustered backgrounds, which are challenging in salient object detection.

SOD contains 300 images from Berkeley segmentation BSD300 [31] dataset. This dataset includes many images with complex natural scenes making it challenging as well.

As in [17] we train our saliency model using the training set that contains 2500 images in MSRA-B dataset, split by [17]. The trained model is then tested on all datasets.

5.2. Evaluation criteria

We adopt the most widely used precision-recall (PR) criterion, where saliency maps, normalized into the range of $[0, 255]$, are binarized at each integer in the range of $[0, 255]$ for computing precision and recall values, which are averaged over all images in each dataset for saliency evaluation. For quantitative evaluation, we also compute the maximal F-measure of the average precision-recall curve, as in [4]. F-measure is a harmonic mean of precision and recall, and is defined as $F_\beta = \frac{(1+\beta^2) \times \text{precision} \times \text{recall}}{\beta^2 \text{precision} + \text{recall}}$, where β^2 is set to 0.3 to emphasize on precision as previous works [1, 4, 41].

Furthermore, we also use the recently suggested mean absolute error (MAE) criterion [33] between a continuous

saliency map \mathcal{S} and the binary ground truth \mathcal{G} for all pixels x_i , *i.e.*,

$$MAE = \frac{1}{|\mathcal{S}|} \sum_i |\mathcal{S}(x_i) - \mathcal{G}(x_i)| \quad (9)$$

where $|\mathcal{S}|$ denotes the number of pixels in the image.

5.3. Performance analysis for the proposed model

To analyze different configurations of the proposed approach as well as different baselines, we plot their generated PR curves on the three benchmark datasets in Figure 4 and show the F_β scores and MAEs in Table 1. As one of the baselines, we consider using the traditional features (including color, texture and regional property) with our compact feature integration scheme, but without HARF representation. As can be seen from the results, with the use of the HARF₁ or HARF₂ description over the traditional features, the PR curves on the three datasets are consistently elevated, thus F_β scores increase correspondingly (*e.g.*, with HARF₂ description, the F_β on SOD dataset increases from 0.699 to 0.714). In Table 1, we can further observe the decrease of MAEs thanks to the use of HARF (*e.g.*, with HARF₁ representation, the MAE on MSRA-B decreases from 0.092 to 0.073).

We also validate the saliency performance for the combination of CNN and traditional features. In this case we observe that a higher performance is achieved by combining the two types of features, compared to using one of them only. With the HARF₁ or HARF₂ representation over the CNN and traditional features, the performance of salient object detection increases even further, yielding a substantial improvement. Overall, compared to using traditional features only, the full HARF₁ (HARF₂) description decreases MAE by 20.6% (22.8%) and improves F_β score by 5.6% (4.6%) on average. The saliency maps generated for the example image by using different configurations of the proposed approach are shown in Figure 5. Clearly, the quality of saliency maps is gradually improved as more components are integrated.

To explore if the last two fully-connected layers of CNN (which carry more semantic information than the lower ones) contribute to salient object detection or not, we generate saliency results by using only the lower CNN layers (layer 1 to 5) as well as all CNN layers (layer 1 to 7). From Figure 4 and Table 1, we can directly see that using all CNN layers achieves higher performance than using the lower ones only, which validates the proposed CNN feature description for saliency detection.

To analyze the sensitivity of HARF construction with respect to the number of super-regions in the segmentation tree (parameter β), we compute the saliency performance for increasing values of β . As shown in Figure 6, performance is improved when β increases and almost reaches saturation at value 7.

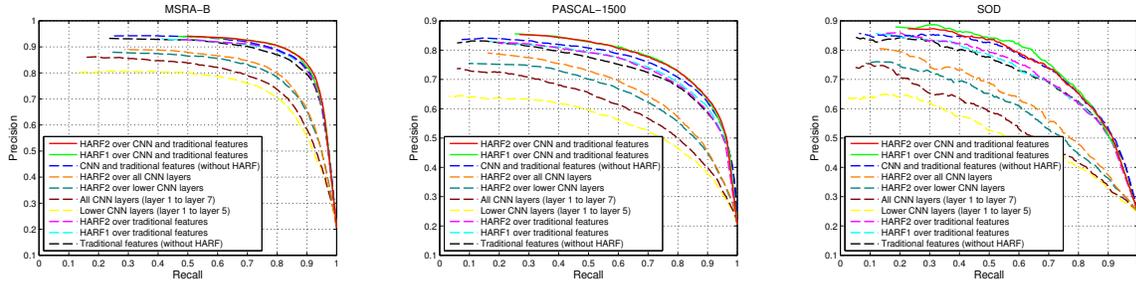


Figure 4. PR curves of the proposed model using different configurations on MSRA-B, PASCAL-1500 and SOD datasets.

Configuration	MSRA-B		PASCAL-1500		SOD	
	F_β	MAE	F_β	MAE	F_β	MAE
Traditional features (without HARF)	0.853	0.092	0.718	0.141	0.699	0.180
HARF ₁ over traditional features	0.864	0.073	0.732	0.121	0.706	0.162
HARF ₂ over traditional features	0.865	0.072	0.730	0.121	0.714	0.159
Lower CNN layers (layer 1 to layer 5)	0.747	0.211	0.573	0.295	0.531	0.321
All CNN layers (layer 1 to layer 7)	0.768	0.193	0.612	0.279	0.569	0.308
HARF ₂ over lower CNN layers	0.799	0.134	0.653	0.212	0.611	0.255
HARF ₂ over all CNN layers	0.811	0.130	0.672	0.209	0.634	0.246
CNN and traditional features (without HARF)	0.868	0.085	0.745	0.131	0.735	0.167
HARF ₁ over CNN and traditional features	0.878	0.067	0.760	0.112	0.759	0.149
HARF ₂ over CNN and traditional features	0.879	0.064	0.759	0.109	0.737	0.146

Table 1. F_β (higher is better) and MAEs (smaller is better) of the proposed model using different configurations on the three datasets.

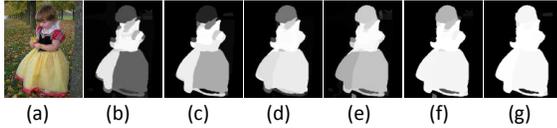


Figure 5. Saliency maps generated by the proposed approach using different configurations: (b) traditional features, (c) HARF₁ over traditional features, (d) HARF₂ over traditional features, (e) CNN and traditional features, (f) HARF₁ over CNN and traditional features, (g) HARF₂ over CNN and traditional features.

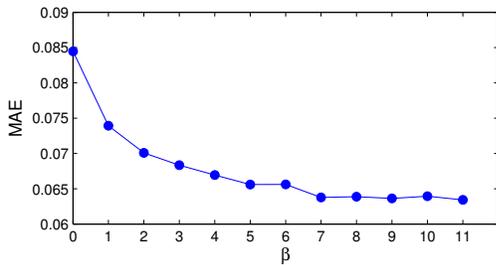


Figure 6. MAE on MSRA-B dataset by varying β for HARF₂ construction.

Run-time: Our current unoptimized MATLAB code for the proposed model on a PC with a GTX Titan X GPU (for CNN feature computation) and an Intel i5 3.2GHz CPU (for all other components) takes around 30h for training and 22s for testing on an image of resolution 400×300 . The most time-consuming component is gPb region segmentation, which takes 17s on CPU while setting the resizing factor for eigenvector computation to 0.5. However, it can be significantly accelerated by use of GPU computing [6].

5.4. Comparison to the state-of-the-art models

For performance comparison, the proposed approach is compared to the *top 6 models* ranked in the recent benchmark report [4], including DRFI [17], RBD [43], DSR [24], MC [15], HDCT [20] and HS [40], and ST model [28] which is not covered in [4]. All the salient object detection results of the compared models are generated using the authors' implementations with their default parameters.

5.4.1 Quantitative comparison

The saliency performance of different models on the three datasets is shown in Figure 7 for PR curves and in Table 2 for F_β scores and MAEs. Obviously, all models achieve considerable higher performance on MSRA-B dataset, compared to the experiments on the other two more challenging datasets. Moreover, the proposed approach whether using HARF₁ or HARF₂ consistently outperforms previous models in terms of all criteria on each dataset. Notice that the PR curves of the proposed model are always higher than others at the top-right corners, which suggests that our model highlights salient objects in a more complete manner thanks to the HARF representation. Quantitatively, compared to the best of the reference models, our proposed model with HARF₁ (HARF₂) description *increases* F_β score by 4.9% (3.94%) and *decreases* MAE by 31% (32.8%) on average, which is a very significant improvement.

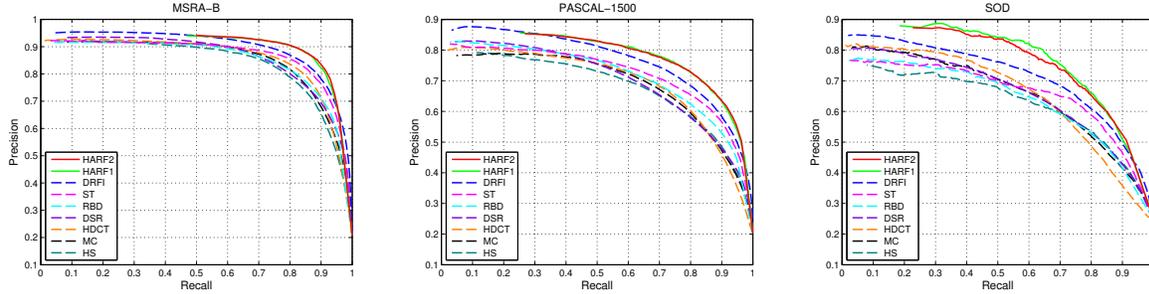


Figure 7. PR curves of different models on MSRA-B (left), PASCAL-1500 (middle) and SOD (right) datasets.

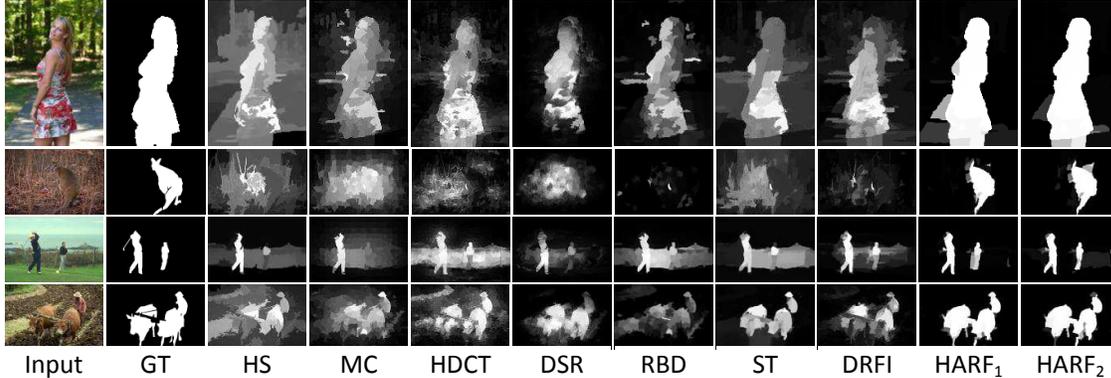


Figure 8. Saliency maps generated by seven state-of-the-art models, and by the proposed model with HARF₁ and HARF₂ respectively.

Model	MSRA-B		PASCAL-1500		SOD	
	F_β	MAE	F_β	MAE	F_β	MAE
HS	0.814	0.161	0.674	0.224	0.628	0.273
MC	0.827	0.144	0.690	0.194	0.649	0.244
HDCT	0.833	0.152	0.695	0.190	0.661	0.226
DSR	0.820	0.116	0.683	0.163	0.648	0.213
RBD	0.821	0.110	0.697	0.154	0.639	0.207
ST	0.844	0.129	0.708	0.178	0.663	0.229
DRFI	0.855	0.119	0.735	0.158	0.695	0.198
HARF ₁	0.878	0.067	0.760	0.112	0.759	0.149
HARF ₂	0.879	0.064	0.759	0.109	0.737	0.146

Table 2. F_β scores (higher is better) and MAEs (smaller is better) of different models on the three benchmark datasets.

5.4.2 Qualitative comparison

Figure 8 shows some saliency maps generated by different models. These clearly demonstrate that the proposed model outperforms other state-of-the-art models not only quantitatively but also qualitatively. Based on these examples we make the following observations:

Heterogeneous appearances: For some images containing salient objects with heterogeneous appearances (*e.g.*, the first row), previous models tend to highlight only part of salient regions, whereas the proposed model with HARF₁ or HARF₂ representation is able to pop out the whole salient object from background regions.

Low contrast: For some images showing low contrast

to salient objects (*e.g.*, the 2nd row), the proposed model can suppress irrelevant background regions and can highlight salient objects that have more well-preserved boundaries compared to the other reference models.

Multiple salient objects: The proposed model generates better looking saliency maps for those images containing multiple salient objects (*e.g.*, the last two rows).

6. Conclusion

We proposed a novel hierarchy-associated rich features (HARF) construction framework, which incorporates various elementary features from both the target region to be described and super-regions in a segmentation hierarchy. The HARF structure is able to represent the global context of the whole salient object or background, which allows much more robust and discriminative features. To enrich the elementary features for HARF construction, multi-layered deep learning features are introduced to complement traditional features. With regional contrast evaluation, all these elementary features are effectively incorporated into the proposed HARF framework. Based on the HARF representation, regional saliency scores are estimated through a boosted predictor. Experiments on the most widely used and challenging benchmark datasets demonstrated that the proposed approach substantially outperforms the state-of-the-art salient object detection models both quantitatively (over 31% MAE decrease) and qualitatively.

References

- [1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *CVPR*, 2009. 2, 6
- [2] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *TPAMI*, 33(5):898–916, 2011. 3
- [3] A. Borji, M.-M. Cheng, H. Jiang, and J. Li. Salient object detection: A survey. *arXiv preprint*, 2014. 2
- [4] A. Borji, M.-M. Cheng, H. Jiang, and J. Li. Salient object detection: A benchmark. *ArXiv e-prints*, 2015. 6, 7
- [5] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *TPAMI*, 35(1):185–207, 2013. 2
- [6] B. Catanzaro, B.-Y. Su, N. Sundaram, Y. Lee, M. Murphy, and K. Keutzer. Efficient, high-quality image contour detection. In *ICCV*, 2009. 7
- [7] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai. Fusing generic objectness and visual saliency for salient object detection. In *ICCV*, 2011. 2
- [8] M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Crook. Efficient salient region detection with soft image abstraction. In *ICCV*, 2013. 2
- [9] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu. Global contrast based salient region detection. In *CVPR*, 2011. 2
- [10] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 6
- [11] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. 2014. 4
- [12] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *CVPR*, 2007. 2
- [13] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *TPAMI*, 1998. 2
- [14] Y. Jia and M. Han. Category-independent object-level saliency detection. In *ICCV*, 2013. 2
- [15] B. Jiang, L. Zhang, H. Lu, C. Yang, and M.-H. Yang. Saliency detection via absorbing markov chain. In *ICCV*, 2013. 2, 7
- [16] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li. Automatic salient object segmentation based on context and shape prior. In *BMVC*, 2011. 2
- [17] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li. Salient object detection: A discriminative regional feature integration approach. In *CVPR*, 2013. 2, 5, 6, 7
- [18] P. Jiang, H. Ling, J. Yu, and J. Peng. Salient region detection by ufo: Uniqueness, focusness and objectness. In *ICCV*, 2013. 2
- [19] Z. Jiang and L. S. Davis. Submodular salient region detection. In *CVPR*, 2013. 2
- [20] J. Kim, D. Han, Y.-W. Tai, and J. Kim. Salient region detection via high-dimensional color transform. In *CVPR*, 2014. 2, 7
- [21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 4
- [22] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu. Saliency detection on light field. In *CVPR*, 2014. 2
- [23] X. Li, Y. Li, C. Shen, A. Dick, and A. V. D. Hengel. Contextual hypergraph modeling for salient object detection. In *ICCV*, 2013. 2
- [24] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang. Saliency detection via dense and sparse reconstruction. In *ICCV*, 2013. 2, 7
- [25] Y. Lin, Y. Y. Tang, B. Fang, Z. Shang, Y. Huang, and S. Wang. A visual-attention model using earth mover’s distance-based saliency measurement and nonlinear feature combination. *TPAMI*, 35(2):314–328, 2013. 2
- [26] R. Liu, J. Cao, Z. Lin, and S. Shan. Adaptive partial differential equation learning for visual saliency detection. In *CVPR*, 2014. 2
- [27] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. *TPAMI*, 33(2):353–367, 2011. 2, 6
- [28] Z. Liu, W. Zou, and O. Le Meur. Saliency tree: A novel saliency detection framework. *TIP*, 23(5):1937–1952, May 2014. 2, 7
- [29] S. Lu, V. Mahadevan, and N. Vasconcelos. Learning optimal seeds for diffusion-based salient object detection. In *CVPR*, 2014. 2
- [30] R. Margolin, A. Tal, and L. Zelnik-Manor. What makes a patch distinct? In *CVPR*, 2013. 2
- [31] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, 2001. 6
- [32] V. Movahedi and J. H. Elder. Design and perceptual validation of performance measures for salient object segmentation. In *CVPRW*, 2010. 6
- [33] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, 2012. 2, 6
- [34] C. Scharfenberger, A. Wong, K. Fergani, J. S. Zelek, and D. A. Clausi. Statistical textural distinctiveness for salient region detection in natural images. In *CVPR*, 2013. 2
- [35] X. Shen and Y. Wu. A unified approach to salient object detection via low rank matrix recovery. In *CVPR*, 2012. 2
- [36] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395–1407, 2006. 2
- [37] W. Wang, Y. Wang, Q. Huang, and W. Gao. Measuring visual saliency by site entropy rate. In *CVPR*, 2010. 2
- [38] Y. Wei, F. Wen, W. Zhu, and J. Sun. Geodesic saliency using background priors. In *ECCV*, 2012. 2
- [39] Y. Xie, H. Lu, and M.-H. Yang. Bayesian saliency via low and mid level cues. *TIP*, 34(11):1689–1698, 2013. 2
- [40] Q. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical saliency detection. In *CVPR*, 2013. 2, 7
- [41] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013. 2, 6
- [42] J. Yang and M.-H. Yang. Top-down visual saliency via joint crf and dictionary learning. In *CVPR*, 2012. 2
- [43] W. Zhu, S. Liang, Y. Wei, and J. Sun. Saliency optimization from robust background detection. In *CVPR*, 2014. 2, 7
- [44] W. Zou, K. Kpalma, Z. Liu, and J. Ronsin. Segmentation driven low-rank matrix recovery for saliency detection. In *BMVC*, 2013. 2, 5, 6