Hindawi
Computational Intelligence and Neuroscience
Volume 2021, Article ID 6699335, 13 pages
https://doi.org/10.1155/2021/6699335

Research Article

A Data-Driven and Biologically Inspired Preprocessing Scheme to Improve Visual Object Recognition

Zahra Sadat Shariatmadar and Karim Faez

Electrical Engineering Department, Amirkabir University of Technology, Tehran 15914, Iran

Correspondence should be addressed to Karim Faez; kfaez@aut.ac.ir

Received 14 October 2020; Revised 28 December 2020; Accepted 20 January 2021; Published 29 January 2021

Academic Editor: Akbar S. Namin

Autonomous object recognition in images is one of the most critical topics in security and commercial applications. Due to recent advances in visual neuroscience, the researchers tend to extend biologically plausible schemes to improve the accuracy of object recognition. Preprocessing is one part of the visual recognition system that has received much less attention. In this paper, we propose a new, simple, and biologically inspired pre processing technique by using the data-driven mechanism of visual attention. In this part, the responses of Retinal Ganglion Cells (RGCs) are simulated. After obtaining these responses, an efficient threshold is selected. Then, the points of the raw image with the most information are extracted according to it. Then, the new images with these points are created, and finally, by combining these images with entropy coefficients, the most salient object is located. After extracting appropriate features, the classifier categorizes the initial image into one of the predefined object categories. Our system was evaluated on the Caltech-101 dataset. Experimental results demonstrate the efficacy and effectiveness of this novel method of preprocessing.

1. Introduction

One of the challenges in the field of artificial intelligence is object recognition. The objective of this process is to classify an object into one of the predefined categories. There are various challenges in this field, such as cluttered and noisy background or objects under different illumination and contrast environments. Human beings can detect and classify objects without any effort in a short time. Researchers believe that the recognition system is closer to the human visual system will be better. In other words, numerous studies [1–3] have shown that inspired by the human visual system, the recognition system can be designed with relatively high accuracy. According to the recent advances in visual neuroscience, the researchers tend to develop biologically plausible algorithms to improve the accuracy of the object recognition system. Object recognition considerably relies on image representation, for which, in this paper, a novel biologically inspired model is presented for this stage. Among image representation models, bag-of-words (BoW) representation [4] has been generally employed because it is robust to object scale and translation changes. Three different modules of BoW models are extraction, coding, and pooling of different features. K-means clustering, which is applied for feature coding, will cause severe information loss because of the hard assignment of each feature to the nearest cluster center. So, soft k-means [5] and sparse coding [6] procedures are presented to overcome this problem.

Sparse coding-based methods are broadly used since they have fewer parameters and more reliable performance than soft k-means. Some different sparse coding-related feature coding techniques [6–8] are offered and obtain the best achievement for image presentation. During the sparse coding-based strategies, the image is represented by a vector of sparse codes matching the features in the individual image within the feature pooling module.

In the BoW model, the whole image is the pooling area, and therefore, the spatial information may be lost. Such data can significantly affect recognition accuracy.

A spatial pyramid matching (SPM) plan [9] is introduced to represent an image. This algorithm is divided into finer regions introduced to preserve spatial information.

It can be said that current approaches for object recognition mainly use machine learning methods. There are several solutions for improving the recognition accuracy, such as collecting larger datasets, using powerful learning algorithms, and preventing overfitting by using better techniques. In recent years, significant steps have been taken to make the recognition systems more effective.

Deep Neural Network (DNN) is one of the best algorithms that have shown excellent results on benchmark datasets [10, 11]. One of the most applicable of these networks is Convolutional Neural Networks (CNNs) [12–14]. These networks are variations of multilayer perceptron and are inspired by biological processes. These models have a massive learning capacity in which we can learn about thousands of objects from millions of images by using them.

Another set of methods that are used for object recognition is classifying the salient objects. To the best of our knowledge, the objects usually are more conspicuous than the background in object recognition systems. Recognizing the objects in the human visual system is closely related to saliency detection. The biological systems using this process will be able to remove unneeded information and focus on the essential regions in an image. In this procedure, two factors determine the pertinent information: Top-Down (TD) or Bottom-Up (BU).

There are several ways to extract the highlighted area of the image, including [15–22]. In [15], a new classification scheme is presented which combines CNN and visual attention mechanism; Shariatmadar and Faez [16] proposed a model that combines the bottom-up and top-down features to extract the prominent part of the image; He et al., in [17], extended Itti's model by using structure tensor; Luo et al. [18] identified the salient object based on backbone enhanced network; Yang et al. [19] detected the salient part of the image by introducing the double-random-walks; in [20], the BU and TD features of a single image are used to detect salient region; Wang et al., in [21], proposed a model of saliency detection related to multilevel deep pyramid (MLDP), and finally, in [22], the diver target detection is performed based on a saliency detection method.

Image quality assessment [23], video coding [24, 25], image contrast estimation [26], and image Watermarking [27] are other uses of the salient area extraction.

Researchers have revealed that saliency discovery has an inherent utilization in target recognition [28–31]. In [28], the images are categorized by immediately classifying the obtained saliency maps. Shokoufandeh et al. [29] represent 3D objects by building a hierarchical graph arrangement based on the saliency map. Moosmann et al. [30] used the remarkable characteristics to boost the classifiers for recognizing the objects. In [32], a saliency network in the first layer of the HMAX [33] architecture was used. Frintrop et al. [34] trained the classifier relying on the conspicuous areas instead of the whole image to speed up classification.

In this paper, by mimicking the human visual system and using machine learning algorithms, we designed a system for object recognition. In fact, after simulating the RGCs responses of an RGB image (new representation of the image), the spike map is obtained by selecting an appropriate threshold. The image pixels corresponding to spikes highlight the salient objects (BU saliency detection). Then, the saliency submaps are linearly combined with the entropy coefficients. In the final stage, the features of remarkable objects are extracted and the classifier categorizes it to predefined classes.

Briefly, the central contributions of this research are expressed as follows:

(i) Using the RGCs responses for obtaining a good representation of an image (inspiring human visual system)

(ii) Determining only those pixels generate an action potential (inspiring the spiking neural network in human beings)

(iii) Reducing the computational cost of object recognition by extracting the most salient object in the image

The rest of this paper is prepared as follows. The description of our structure is developed in Section 2. In Section 3, experiments are carried out to assess our proposed system. The discussion and conclusion are given in Sections 4 and 5, respectively.

2. System Overview

Figure 1 shows the general review of the proposed method. At first, the raw image is defined in the CIE LAB color space, and each channel is preprocessed by simulating RGCs in the human retina. Then, the resultant images are fed into a spike generator. In this stage, the pixels which are higher than a predefined threshold are selected. The new images obtained by these pixels are saliency submaps, combined linearly by entropy coefficients, and so the final saliency map (Region of Interest (ROI)) is obtained. Finally, after extracting appropriate features from ROI, the classification step is done, and a label is assigned to each initial image. All of these stages are described in the following sections (since the focus of this paper is to represent an image and to extract ROI, feature extraction and classification procedures are described briefly).

2.1. Image Preprocessing. Image preprocessing is the front-end of each recognition system. In this stage, a good and meaningful representation of the raw image is obtained for further processing in the next steps (in this paper, new description is equivalent to ROI extraction). Various phases of the proposed method for obtaining the new representation of the raw image are as follows.

2.1.1. Color Space Transformation. In this stage, the RGB image is encoded into three color channels defined in the CIE LAB color space. The CIE LAB color is a three-dimensional space that comprises the entire range spectrum of human

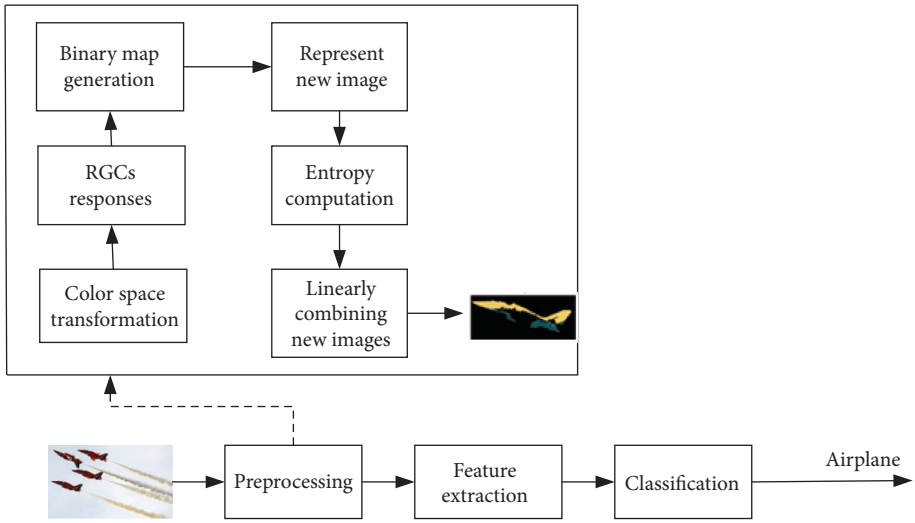

Figure 1: Overall scheme of the proposed method.

color perception. This color space represents color as three states: L (lightness), a (green to red), and b (blue to yellow). Therefore, the four individual colors of human vision are covered (yellow, blue, green, and red). In other words, the LAB color space models are the opponent characteristics of color in the human visual system. Different color spaces (such as CMYK and RGB, which are device-dependent) are not designed to imitate human visual perception. In other words, these spaces model the output of physical devices.

Briefly, in this paper, the aim is to offer a preprocessing method based on the human visual system. Accordingly, LAB color space was selected for subsequent image processing for these reasons: (1) considering the uniform distribution of color in human vision, (2) modelling the opponent properties of color in the human visual system, and (3) imitation of the three-stimulus model in the color vision system of humans by many digital cameras.

For this transformation, two conversions are done: the RGB space to XYZ space and XYZ space to CIELAB space. The formulas of them are shown as follows:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 2.7698 & 1.7517 & 1.1302 \\ 1 & 4.5907 & 0.0601 \\ 0 & 0.0563 & 5.5943 \end{bmatrix} * \begin{bmatrix} R \\ G \\ B \end{bmatrix},$$

$$\begin{cases} L = 116.h\left(\dfrac{Y}{Y_w}\right) - 16, \\ A = 500\left[h\left(\dfrac{X}{X_w}\right) - h\left(\dfrac{Y}{Y_w}\right)\right], \\ B = 200\left[h\left(\dfrac{Y}{Y_w}\right) - h\left(\dfrac{Z}{Z_w}\right)\right], \end{cases} \tag{1}$$

where

$$\begin{aligned} X_w &= 95.04, \\ Y_w &= 100.00, \\ Z_w &= 108.89, \\ h(q) &= \begin{cases} \sqrt[3]{q}, & q \geq 0.008856, \\ 7.787q + \dfrac{16}{116}, & q \leq 0.008856. \end{cases} \end{aligned} \tag{2}$$

2.1.2. Simulating RGCs' Responses. In this stage, we used the same functionality of RGCs in the human retina.

In the human eye, the neurons near the retina's inner surface are referred to as retinal ganglion cells. The visual information is passed to these cells via bipolar and retinal amacrine cells. They differ remarkably in terms of their dimension, associations, and responses to optical stimulation. There is a minimum of five central categories of retinal ganglion cells, which are based on their purposes. One of these classes which are considered in this paper is Midget cells (parvocellular, or P pathway; P cells). Midget retinal ganglion cells map into the parvocellular layers of the lateral geniculate nucleus. Parasol cells are approximately 80% of the total retinal ganglion cells that get information from several rods and cones. They have a fast transfer rate and can react to low-contrast stimuli. They have uncomplicated center-surround receptive fields, where the center can be either OFF or ON while the surrounding is in the opposite mode (the receptive field of a single sensory neuron is the specific area of the sensory space in which a stimulus will activate the firing of the neuron). Three main steps in modelling these cells are as follows (for modelling Midget cells, we further developed receptive fields for peripheral and foveal cells which have ON and OFF parts):

Step 1: initializing basic parameters (standard deviations of the central Gaussians) according to Table 1 [35].

Step 2: making a difference of Gaussian (DOG) filters to model RGC receptive fields.

In this stage, center and surrounding Gaussians of ON and OFF DOG filters for the simulated RGC cells are obtained. Three steps of these filters are as follows:

(i) Making central Gaussian,

$$G = \exp\left(\frac{-x^2}{2\sigma^2}\right). \tag{3}$$

(ii) Converting it to the 2D filter,

$$G = G^T \cdot G. \tag{4}$$

(iii) Normalizing to sum to 1,

$$\text{Gaussian} = \frac{G}{\sum G}. \tag{5}$$

It is emphasized that the goal of this article is not to use the concepts related to spikes and excitatory neurons in different areas of the human brain. The only biological use in this paper is to simulate the response of retinal ganglion cells without spike encoding.

Figure 2 shows the center and surrounding receptive fields for foveal midget cells.

Step 3: applying the RGC models to the image matrix for the foveal pathway.

In this step, the image is convolved to RGC filters. The resultant of this step is six responses: midget foveal/peripheral OFF reactions for each of the three channels L, A, and B. Figure 3 shows the responses of foveal/peripheral OFF midget cells for three channels.

Finally, six maps of a raw image are obtained, which is fed to the next step. We believed that, by getting the responses of RGCs, the better representation of the raw image is achieved.

2.1.3. Binary_Map Generation. In this stage, the front-end of the individual visual system is inspired. This section, which consists of several primary layers of neurons (retina photoreceptors to the primary visual cortex), is shown in Figure 4.

There are two classes of photoreceptors referred to as rods and cones. Rods have a great sensibility to low levels of illumination, and cones require high levels of intensity. These cells that are susceptible to a specific interval of the electromagnetic spectrum convert visual information to neural signals. The outputs of this biological pathway are action potentials.

In physiology, an action potential is related to a short-lasting situation in which the membrane's electrical potential in a cell immediately rises and falls. Action potentials occur in neurons, which are excitable cells. In neurons, cell-to-cell communication is done with action potentials. When the neurons fire, the action potentials or spikes are emitted.

If the image pixels are considered as photoreceptors and the action potential is represented with the binary string, we can emulate the above biological pathway by a linear-nonlinear cascade. A linear function is the linearly combining different weights of bipolar cells, and the nonlinear function is the rectifier function (comparing with a threshold).

This stage aims to generate the binary strings which are obtained by selecting an appropriate threshold. On the contrary, each pixel of RGCs responses is compared to the limit as follows:

$$\text{binary\_Map}_{\text{Res}(i.j)} = \begin{cases} 1, & \text{if Res}(i.j) > \text{Threshold}, \\ 0, & O.W. \end{cases} \tag{6}$$

in which

$$\begin{aligned} Res &= \{\text{Res1.Res2.Res3.Res4.Res5.Res6}\} \\ &= \left\{ \begin{array}{l} P_{\text{foveal-off}_L}.P_{\text{peripheral-off}_L} \\ P_{\text{foveal-off}_A}.P_{\text{peripheral-off}_A} \\ P_{\text{foveal-off}_B}.P_{\text{peripheral-off}_B} \end{array} \right\}. \end{aligned} \tag{7}$$

In the above formula, L, A, and B are three channels of LAB space.

After various experiments, we found that the mean value of each image is the best threshold value for the same image. In other words, the average value of each image is obtained experimentally.

The threshold is suitable when it can detect prominent objects in multiple images. In other words, since the objects in many images are distinguished from the background by the average amount of image pixels, this threshold value seems appropriate (experimental results in the salient object detection (Section 3.3.1)).

If the threshold value is not selected correctly, the preprocessing step will not work successfully. As a result, the object class in the recognition phase is not assigned correctly. In other words, the accuracy of the classification depends on the correct detection of the object in the preprocessing stage. This detection also corresponds to the selection of the threshold. Therefore, with the accurate choice of this measure, the object of the images can be detected with high accuracy, and finally, the correct class of that object in the classification stage can be guaranteed.

2.1.4. New Representation of the Image. So far, we have six binary maps of RGCs which correspond to action potentials of Ganglion cells. Then, by using these maps, those pixels of the raw image in each L, A, and B channel form a new image. So, there will be two new images for each of LAB channels: one for foveal response and one for peripheral response in off-pathway of the ganglion cell:

Table 1: The values of standard deviations for central Gaussian in midget cell.

Cell type	Location	Center std dev (arcmin)	
		On	Off
Midget cells	Foveal	1.4	1.1
	Peripheral	3.3	2.7

(a) (b)

Figure 2: (a) The center and (b) the surrounding receptive fields for foveal midget cells.

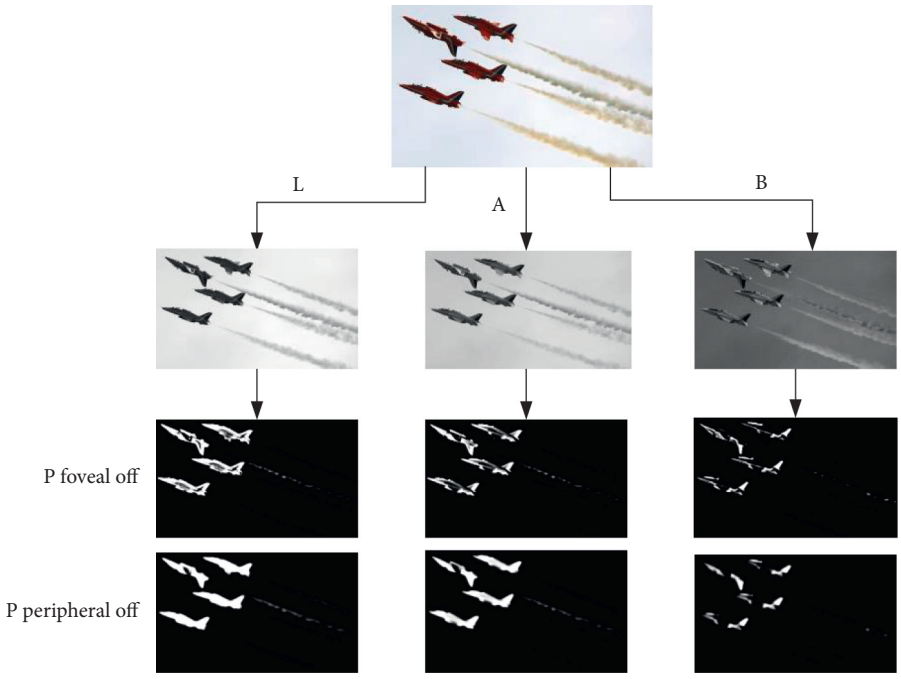

Figure 3: The responses of foveal and peripheral OFF midget cells.

$$\text{New\_Image}_{\text{channel pathway}}(i.j) = \begin{cases} \text{Initial\_Image}_{\text{channel}}(i.j), & \text{if Binary\_Map}_{\text{channel}_{\text{pathway}}}(i.j) = 1, \\ 0, \quad O.W, \end{cases} \tag{8}$$

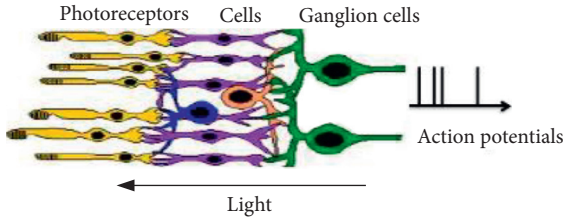

Figure 4: The front-end of the human retina (leading to action potentials).

in which *channel* $\in \{L.A.B\}$ and pathway $\in$ {foveal. peripheral}. At the end of this stage, we have six new images in which their pixels correspond to those neurons that emit an action potential.

2.1.5. ROI Extraction. This stage is the last step in the preprocessing stage which is proposed as follows:

(i) Computing the entropy coefficient from each of the New_Image obtained in the previous step

(ii) Linearly combining New_Images along with entropy coefficients

(iii) Cropping the final result from the image as ROI

The proposed fusion rule for obtained NewImages is as follows:

$$\begin{aligned}\text{Final\_Image} &= \sum_{i=1}^{6} \gamma_i \,\text{New\_Image}_i \\ &= \left(\sum_{i=1}^{3} \alpha_i \left(\text{New\_Image}_{\text{foveal}}\right)_i\right) \\ &\quad + \left(\sum_{j=1}^{3} \beta_j \left(\text{New\_Image}_{\text{peripheral}}\right)_j\right).\end{aligned} \tag{9}$$

In this equation, the final image is formed by combining $\text{New\_Image}_{\text{foveal}}$ with α_i ($i = 1, 2, 3$) and $\text{New\_Image}_{\text{peripheral}}$ with β_j ($j = 1,2,3$). The β and α coefficients (γ coefficients) are determined by entropy measure, and i and j correspond to L, A, and B channels. The texture of the image is specified by a statistical measure of randomness called entropy. In each image, this criterion is determined as

$$E(S) = -\sum_{i=1}^{m} p_i \log\, p_i. \tag{10}$$

In the above equation, m is the maximum image value, p_i is the probability of value i, and γ coefficients are defined as follows:

$$\begin{aligned}\gamma_i &= \frac{\lambda_i}{\sum_{j=1}^{6} \lambda_j}, \\ \lambda_j &= \frac{1}{E\left(I_j\right)},\end{aligned} \tag{11}$$

where $j = 1, \ldots, 6$ representing the index of new images obtained by two pathways of L, A, and B channels. $E(I_j)$ is the entropy of the jth and γ_i is the weight of the ith New_Image.

At the end of this stage, Final_Image is cropped and then fed to the feature extraction stage.

The steps of image preprocessing are summarized in Algorithm 1.

2.2. Feature Extraction and Classification. Feature Extraction is one of the most crucial stages of visual recognition. In this stage, the discriminant features for better classification in the next processing are extracted. In this paper, we used the Log-Gabor function [36] for obtaining localized frequency information. Log-Gabor filters give useful information on receptive fields from the simple cell located in V1 of the human brain. For exploiting the diversity of shape characteristics of an image, a bank of Log-Gabor filters was used (at three scales and eight orientations). Then, we used the Principle Component Analysis (PCA) for dimensionality reduction. The obtained vectors are 128-dimensional. After extracting PCA vectors of different photos, Support Vector Machine (SVM) is trained and used as a classifier. In this paper, a simple, linear, and multiclass SVM is applied. Here, binary classifiers are created, and a distinction is made between one label and the rest of the labels (one-versus-all). When a new instance has come for classification in this case, a winner-takes-all procedure is done.

This paper used the SVM with the linear kernel because its structure can be efficiently implemented in the cortex [37]. Also, the user must select a few parameters and does not need to specify parameters for different kernels.

Also, the experiments using nonlinear kernels were performed. These kernels did not significantly improve the proposed model. Therefore, a linear kernel seems to be the right choice. In this paper, the linear Lib-SVM classifier [38] is used.

3. Experimental Results

Our proposed method is compared with other state-of-the-art algorithms in both saliency detection and object recognition in this section. In other words, the output of the preprocessing stage in the explained scheme is considered as a salient object. Then, the extracted remarkable object is classified as a predefined class (Table 2).

3.1. Experimental Setup

(1) *Datasets.* In the salient object detection stage (the output of preprocessing unit), the proposed algorithm is evaluated on two publicly datasets: MSRA-B [39] and ECSSD [40], and in the classification phase, the Caltech-101 [41] database is used. MSRA-B has many natural images (so, the comparison is made on a large scale), and ECSSD has structurally complex images (Table 2). The ground truth in these two datasets has segmented manually.

Input: raw image (I)
Output: saliency region (S)
Step 1: RGB to CIE LAB color space conversion:

(1) $\mathbf{I_{XYZ}} \leftarrow \mathbf{I_{RGB}}$
RGB space is transformed to XYZ space according to formula (1)
(2) $\mathbf{I_{LAB}} \leftarrow \mathbf{I_{XYZ}}$
XYZ space is transformed to LAB space according to formula (2)

Step 2: RGC response's simulation:

(1) $\left\{\mathbf{Responses}_{(\mathbf{off})}^{(\mathbf{foveal.\ peripheral})}\right\}_1^6 \leftarrow \{\mathbf{Channel}\}_{\mathbf{L.A.B}} * \mathbf{Do\ G\ filter}$
The L, A, and B channels are convolved with the Gaussian function based on the midget cells' in the human retina (the standard deviation of this Gaussian function is considered according to Table 1)

Step 3: binary map generation

(1) $\mathbf{Binary\ map} \leftarrow \mathbf{Threshol\ d}\ (\{\mathbf{Responses}\}_1^6)$
By selecting an appropriate threshold, the Responses are converted to binary images (the threshold is chosen based on the mean value of each gray image)

Step 4: the new image representation

(1) $\{\mathbf{New\ Image}\}_{\mathbf{L.A.B}}^{\mathbf{foveal.peripheral}} \leftarrow \mathbf{Mask}\{\mathbf{Initial\ image}\}$
The mask operation is done by using the Binary maps in the previous stage

Step 5: ROI extraction

(1) $\mathbf{S} \leftarrow \boldsymbol{\alpha}\{\mathbf{New\ image_{foveal}}\}_{\mathbf{off}} + \boldsymbol{\beta}\{\mathbf{New\ Image_{peripheral}}\}_{\mathbf{off}}$
All images are combined by using entropy quantity according to formula (9)
The $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ coefficients are calculated according to formula (10)

Algorithm 1: The steps of Image preprocessing.

Table 2: Various specifications of three datasets.

Dataset	Ref.	# of samples	Number of objects	Type of background	Scene resolution
MSRA	[39]	5000	~1	Simple	~400∗300
ECSSD	[40]	1000	~1	Complex	~400∗300
Caltech-101	[41]	9146 (101 object categories)	~1	~simple	~300∗200

(2) *Implementation details.* In our experiment, our scheme is carried out in MATLAB 2013b on a Dell Vostro 3300 with an Intel i5 M520 2.4 GHz CPU and 8 GB RAM.

3.2. Evaluation Metrics. In our experiments, a widely selected metric called F-measure is used. This measure is calculated to assess the performance of schemes comprehensively and determined as follows:

$$F_\gamma = \frac{(1+\gamma)\text{Precision} \times \text{Recall}}{\gamma \times \text{Precision} + \text{Recall}}. \tag{12}$$

In the above equation, the precision and recall quantities are defined as follows. γ is fixed to 0.3 as proposed by [42] to highlight precision:

$$\text{Precision} = \frac{B \cap G}{B}, \qquad \text{Recall} = \frac{B \cap G}{G}, \tag{13}$$

where G is the ground truth in the above relation and B is the binary map of the salient object.

3.3. Performance Evaluation

3.3.1. Salient Object Detection. In this section, the proposed method is compared with various saliency detection schemes, which include GC [43], MC [44], HC [45], RC [45], ST [46], SF [47], HS [40], RBD [48], HDCT [49], DRFI [50], GMR [51], DSR [52], and MFS [53].

In Figure 5, the maximum F-measure using an adjusted threshold is used to compare various methods on the two benchmark datasets. According to this figure, our proposed method works on par with the most reliable schemes on the two databases. Notably, for the ECSSD and MSRA datasets, our method's F-measure is only 3.39% and 3.79% less than the best model [50], respectively. Compared with all the other schemes, our approach is slightly worse than the best systems on these datasets.

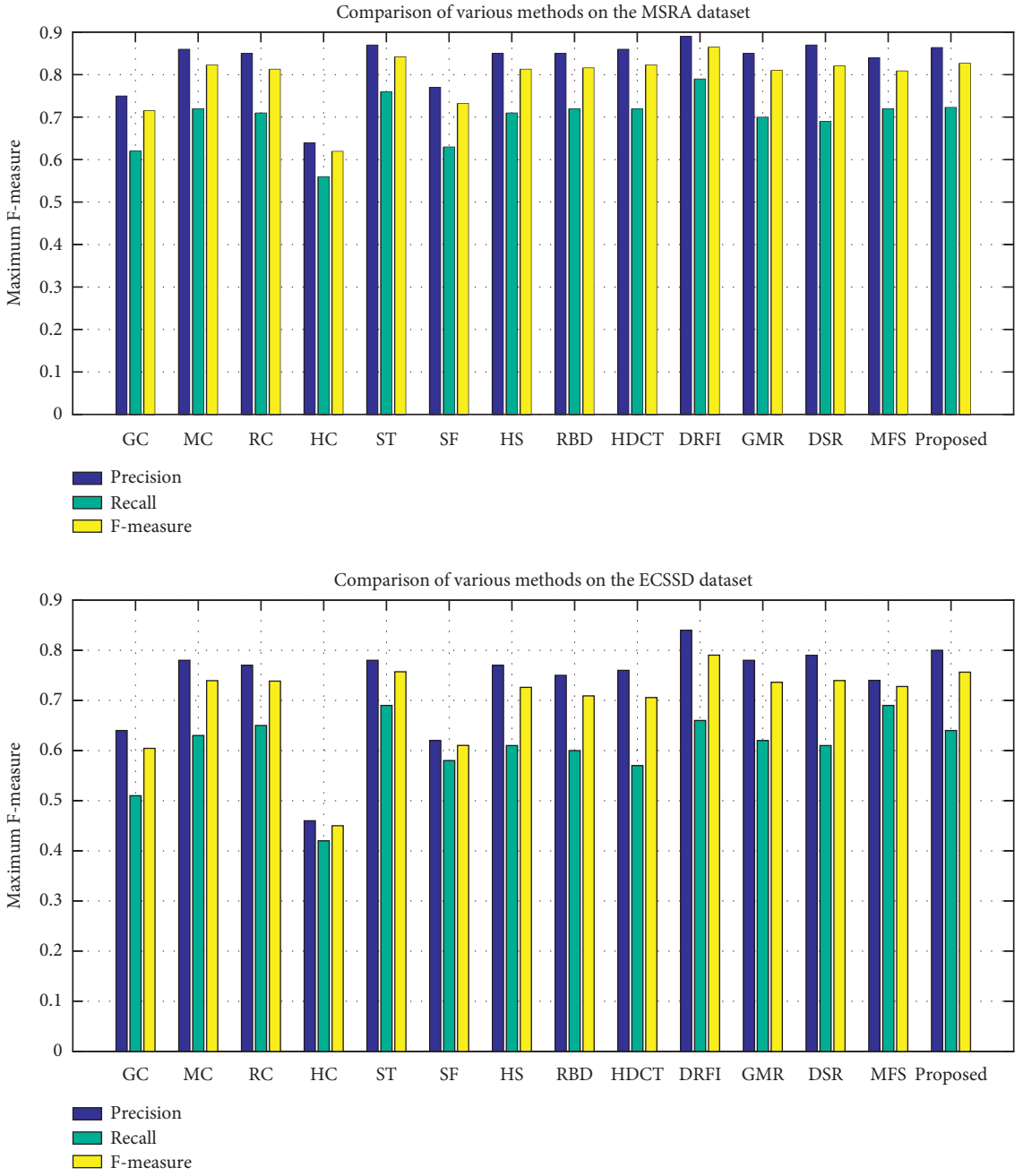

Figure 5: Quantitative comparison of our method with different methods on two widely used datasets MSRA-B and ECSSD in terms of maximum F-measure.

3.3.2. Object Recognition. As mentioned in Section 3.1, we used the Caltech-101 database [41] to evaluate our proposed method's performance. This database includes 101 object classes and each of which comprises between 40 and 800 images. After considering different training images, we found experimentally that 15 training images for each class are the best. We used 15 images for training, and 50 other chosen images for testing in each group (pictures are selected randomly). All the algorithms compared with the proposed method in the next section have also selected their training and testing images in the same way. If less than 65 images were available for a category, we trained on 15 random pictures per class and then examined on all unused images. For describing the results among this database, we estimate the mediocre efficiency for every category. Table 3 shows our per-class results. This table exhibits the first ten best-classified categories utilizing the Caltech-101 database.

Typical images from these classes are shown in Figure 6.

Following the images in Figure 6, it can be said that our results are quite good on average because

Table 3: The first ten best-classified categories utilizing the Caltech-101 database.

Category	Motorbikes	Helicopter	Wrench	Revolver	Saxophone	Inline skate	Scorpion	Menorah	Mandolin	Windsor chair
Accuracy (%)	95	93	93	91	91	90	90	90	89	88

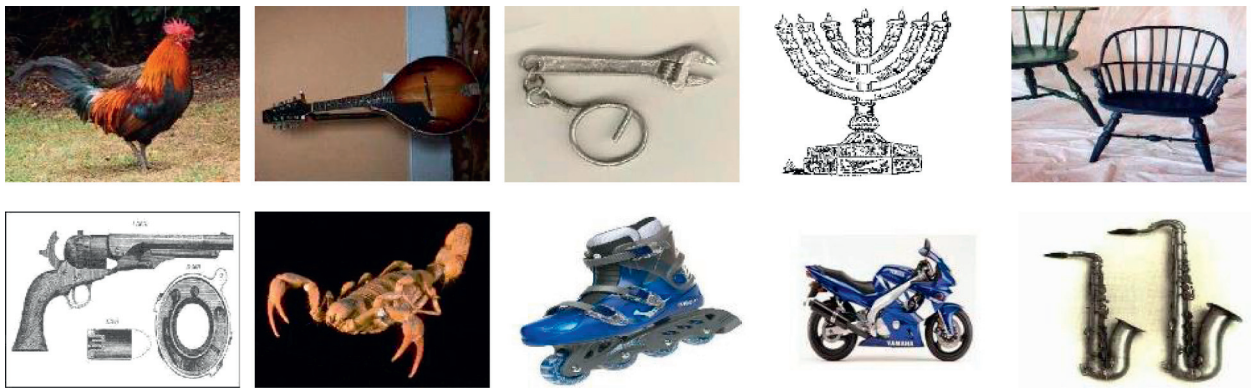

Figure 6: Several typical images from Caltech-101 dataset.

(i) The pictures of these classes have a simple background, and the proposed method in ROI extraction can obtain a good representation of an object.

(ii) In these classes, the shape is an excellent discriminatory feature, and so our scheme in saliency detection is quite good.

(iii) In categories such as wild-cat, hedgehog, and butterfly, our algorithm does not have acceptable accuracy. The reasons for this observation are the cluttered background and lack of appropriate features. On the contrary, in the images with natural objects, additional features such as color and texture should be considered. It seems that, by combining different elements efficiently, we can appropriately recognize the objects of these categories.

Generally, the objects are divided into two categories (natural and human-made). Physical objects usually have a cluttered background and various colors, so our proposed method in recognizing these objects is impotent (of course, if the background is simple, our algorithm will be relatively good).

In the following, the comparison between our method and other algorithms are given in Table 4. This comparison is accomplished in terms of eight categories of the Caltech-101 dataset. It should be noted that, for a fair comparison, the training and testing images must be the same for all methods. Therefore, the listed results of other schemes are not used (different algorithms were reimplemented by the authors), and the accuracy values are reported using the same training and testing images for all algorithms.

The scheme developed by Berg [54] is one of the most reliable nonbiologically inspired methods. In his work, the shapes are presented by sampling several pixel positions obtained from the output of an edge discovering process. Lazebnik et al. [9] adopted the kernels of spatial pyramid matching of visual words of Scale Invariant Feature Transform (SIFT) [55] descriptors. HMAX [33], a famous architecture for extracting the biologically inspired visual features, achieves 51.2 percent accuracy when used with SVM as a classifier (in their scheme, 15 images are used for training). This paradigm, based on considerations of visual receptive fields discovered in monkey and cat visual cortex, is usually extremely slow for real-time purposes. Also, the researchers in [56] suggest a neuromorphic visual object recognition structure motivated by neuroscience principles of recognition and visual attention in the human brain. Lu et al., in [57], enhanced the HMAX model by modifying the path selection in S2 layer of it. For this purpose, they used the concept of salient region in any image. Finally, Norizadeh et al. [58] developed an enhanced model of HMAX based on SIFT features. These characteristics are used to select the regions with most information.

Machine learning schemes develop a model based on example data, identified as "training data" to make decisions. According to this definition, the methods described above all use machine learning algorithms. According to Table 4, different methods in various classes can have the best accuracy. Some of the specifications of these methods are summarized in Table 5.

As reported in Table 4, it seems that our proposed method works better than other research studies and is 2% less than the Khosla model in faces objects. It can be said that the new representation of images by using the RGCs responses can obtain an accurate ROI in pictures by highlighting the salient objects and removing the redundant information.

Also, MATLAB code of some methods, such as [33], are publicly available, which has been used in our comparisons. However, some schemes, such as [55–57], have been reimplemented by the authors.

Table 4: Recognition accuracy for different schemes in eight categories of Caltech-101 dataset.

Method	Accuracy (%)							
	Airplanes	Chair	Faces	Motorbikes	Rooster	Helicopter	Wrench	Saxophone
Berg [54]	74	71	70	76	72	73	71	69
Lazebnik et al. [9]	88	84	89	90	83	87	87	82
HMAX [33]	86	82	87	87	79	84	83	81
Khosla et al. [55]	90	87	**92**	91	85	91	90	89
Lu et al. [56]	89	90	87	88	86	87	89	88
Norizadeh et al. [57]	88	91	89	92	84	90	91	86
Proposed method	**92**	**94**	90	**95**	**87**	**93**	**93**	**91**

Table 5: Specifications of various methods.

Method	Type of model	Learning	Classifier
Berg [54]	Nonbiologically inspired	Supervised	K-NN
Lazebnik et al. [9]	Nonbiologically inspired	Supervised	SVM
HMAX [33]	Biologically inspired	Supervised	SVM
Khosla et al. [55]	Biologically inspired	Supervised	K-NN + SVM
Lu et al. [56]	Biologically inspired	Supervised	SVM
Norizadeh et al. [57]	Biologically inspired	Supervised	SVM
Proposed	Biologically inspired	Supervised	SVM

It should be noted that, in this paper, our focus is on the preprocessing stage. The goal is to show that if image preprocessing is done with high accuracy, the classification precision will eventually increase. In other words, the goal is to find an efficient shallow model that can achieve high accuracy in classification by performing precise preprocessing. Also, the amount of training data in this issue is much less than the data available for deep network training. For example, the amount of data in Caltech is much smaller than in ImageNet. So, we used the Caltech-101 dataset to compare shallow networks. For this reason, the findings were not compared to the CNN networks.

4. Discussion

Since the proposed method uses the RGCs responses in the human retina for image presentation, it achieved the best performance over other methods in the Caltech-101 dataset. On the contrary, if we have an appropriate representation of an image in the initial stages of the recognition system, we will send salient information for the classifier in the last step. In the human retina, the essential information and the less critical data are transferred into the visual cortex. By using this idea, we obtain a new representation of the raw image with RGCs responses.

Some of the capabilities of our proposed approach include

(i) If the image has more than one object, our proposed method can discover and eliminate unrelated data from an image (if the background is not too cluttered)

(ii) In human-made objects and simple background of images, our algorithm can extract the most salient information

(iii) Simplicity and robustness to illumination and noise are other advantages of our scheme

Some of the limitations of our proposed approach include

(i) In our algorithm, only the shape features are used. Integration features can be used to obtain better recognition accuracy. For example, the color feature can be used in the preprocessing step (different channels of color (Red-off, Red-on, Blue-off, and so on) according to visual perception in the human retina) or the feature extraction stage (for example, color naming).

(ii) The proposed method does not work relatively well, mainly in dealing with images with crowded backgrounds (the results of salient object detection for ECSSD database in Section 3.3.1). A raw idea to solve this problem is to consider a combination of different channels in various color spaces to use the unique color information of the objects. Acceptance of this idea depends on further experiments.

(iii) Another limitation of our method is that we do not consider the occlusion challenge. On the contrary, our proposed method can work well when the objects are without any occlusion. Of course, the suggested algorithm may work well on some partially occluded images, but if there is a lot of occlusion in the image, our scheme will not be usable.

(iv) Matlab software is used for implementing our method. The proposed algorithm's speed can be improved by using the C++-based execution or employing the parallelization techniques.

5. Conclusion

In this paper, a visual recognition scheme using natural scenes is presented. In this study, we focus on the preprocessing stage and reveal that the image is more similar to

the processed image of the human retina and the final classification accuracy would be higher. At first, each raw image is represented in six activation maps, which are the simulated reactions of retinal ganglion cells of human retina (two responses of Midget cells in three channels of LAB color space). After modelling the RGC's response, an acceptable threshold is selected and a Binary_Map is created by comparing the RGC's response to that threshold. This binary mask forms a new representation of the raw image. Finally, by computing the entropy coefficients of each image and combining them, the final image is obtained. This image highlights the most salient information and removes the redundant background.

The results of various experiments presented in Section 3 illustrate the suitability of our proposed methods for recognizing objects. Some of the future works for obtaining better recognition accuracy can be mentioned as follows:

(i) Using different color channels by inspiring color perception in the human retina

(ii) Integrating shape features with texture and color

(iii) Implementing coarse and fine classification in the last step of the visual recognition system: this type of classification is speculated to arise in the inferior temporal cortex (IT) in the human brain [59]

(iv) Using the PASCAL-S and Judd datasets for considering the images with complex scenes

(v) Investigating the proposed preprocessing approach for deep-structured networks (testing the proposed method on the image net database) and achieving reasonable results for fine-grained classification

Data Availability

The data used to support the findings of this study are included within the supplementary information file and are publicly available.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors acknowledge the Amirkabir University of Technology (Tehran Poly Technique) for financially supporting and processing equipment of this work.

Supplementary Materials

The three public datasets used to support this study are available at , , and . (*Supplementary Materials*)

References

[1] T. Serre, L. Wolf, and T. Poggio, "Object recognition with features inspired by visual cortex," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA, June 2005.

[2] M. Ghodrati, S. M. Khaligh-Razavi, R. Ebrahimpour, K. Rajaei, and M. Pooyan, "How can selection of biologically inspired features improve the performance of a robust object recognition model?," *PLoS One*, vol. 7, no. 2, 2012.

[3] H.-Z. Zhang, Y.-F. Lu, T.-K. Kang, and M.-T. Lim, "B-HMAX: a fast binary biologically inspired model for object recognition," *Neurocomputing*, vol. 218, pp. 242–250, 2016.

[4] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *Proceedings of the IEEE Conference on Computer Vision*, pp. 1470–1477, Nice, France, October 2003.

[5] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, Minneapolis, MN, USA, June 2007.

[6] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1794–1801, Miami, FL, USA, June 2009.

[7] S. Gao, I. W.-H. Tsang, and L.-T. Chia, "Laplacian sparse coding, hypergraph Laplacian sparse coding, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, 2013.

[8] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Learning locality constraint linear coding for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3360–3367, San Francisco, CA, USA, June 2010.

[9] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2169–2178, New York, NY, USA, June 2006.

[10] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems (NIPS)*, vol. 25, no. 2, 2012.

[11] O. Russakovsky, J. Deng, H. Su et al., *ImageNet Large Scale Visual Recognition Challenge*, arXiv:1409.0575, 2014.

[12] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 609–616, Montreal, Quebec, Canada, June 2009.

[13] N. Pinto, D. Doukhan, J. J. DiCarlo, and D. D. Cox, "A high-throughput screening approach to discovering good forms of biologically inspired visual representation," *PLoS Computational Biology*, vol. 5, no. 11, 2009.

[14] S. C. Turaga, J. F. Murray, V. Jain et al., "Convolutional networks can learn to generate affinity graphs for image segmentation," *Neural Computation*, vol. 22, no. 2, pp. 511–538, 2010.

[15] N. Li, X. Zhao, Y. Yang, and X. Zou, "Objects Classification by Learning-Based Visual Saliency Model and Convolutional Neural Network," *Computational Intelligence and Neuroscience*, vol. 2016, Article ID 7942501, 2016.

[16] Z. S. Shariatmadar and K. Faez, "Visual saliency detection via integrating bottom-up and top-down information," *Optik*, vol. 178, pp. 1195–1207, 2019.

[17] Z. He, X. Chen, and L. Sun, "Saliency Mapping Enhanced by Structure Tensor," *Computational Intelligence and Neuroscience*, vol. 2015, Article ID 875735, 2015.

[18] R. Luo, H. Huang, and W. Wu, "Salient object detection based on backbone enhanced network," *Image Vision and Computing*, vol. 95, 2020.

[19] J. Yang, X. Fang, L. Zhang, H. Lu, and G. Wei, "Salient object detection via double random walks with dual restarts," *Image Vision and Computing*, vol. 93, 2020.

[20] H. Tian, Y. Fang, Y. Zhao, W. Lin, R. Ni, and Z. Zhu, "Salient region detection by fusing bottom-up and top-down features extracted from a single image," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4389–4398, 2014.

[21] H. Wang, L. Dai, Y. Cai, L. Chen, and Y. Zhang, "Saliency detection by multilevel deep pyramid model," *Journal of Sensors*, vol. 2018, Article ID 8249180, 2018.

[22] J. Zhu, S. Yu, L. Gao, Z. Han, and Y. Tang, "Saliency-based diver target detection and localization method," *Mathematical Problems in Engineering*, vol. 2020, Article ID 3186834, 2020.

[23] K. Gu, S. Wang, H. Yang et al., "Saliency-guided quality assessment of screen content images," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1098–1110, 2016.

[24] H. Hadizadeh and I. V. Bajic, "Saliency-aware video compression," *IEEE Transactions on Image Processing*, vol. 23, no. 1, pp. 19–33, 2014.

[25] Y. Li and Y. Li, "A fast and efficient saliency detection model in video compressed-domain for human fixations prediction," *Multimedia Tools and Applications*, vol. 76, no. 24, pp. 26273–26295, 2017.

[26] K. Gu, W. Lin, G. Zhai, X. Yang, W. Zhang, and C. W. Chen, "No-reference quality metric of contrast-distorted images based on information maximization," *IEEE Transactions on Cybernetics*, vol. 47, no. 12, pp. 4559–4565, 2017.

[27] Y. Zhang and Y. Sun, "An image watermarking method based on visual saliency and contourlet Transform," *Optik*, vol. 186, pp. 379–389, 2019.

[28] D. Walther, U. Rutishauser, C. Koch, and P. Perona, "On the usefulness of attention for object recognition," in *Workshop on Attention and Performance in Computational Vision at ECCV*, pp. 96–103, Prague, Czech Republic, 2004.

[29] A. Shokoufandeh, I. Marsic, and S. J. Dickinson, "View-based object recognition using saliency maps," *Image and Vision Computing*, vol. 17, no. 5-6, pp. 445–460, 1999.

[30] F. Moosmann, D. Larlus, and F. Jurie, "Learning saliency maps for object categorization," in *Proceedings of the ECCV'06 Workshop on the Representation and Use of Prior Knowledge in Vision*, Graz, Austria, May 2006.

[31] B. Jiang, J. Yang, Z. Lv, and H. Song, "Wearable vision assistance system based on binocular sensors for visually impaired users," *IEEE Internet of Things journal*, vol. 6, no. 2, pp. 1375–1383, 2018.

[32] S. Han and N. Vasconcelos, "Biologically plausible saliency mechanisms improve feedforward object recognition," *Vision Research*, vol. 50, no. 22, pp. 2295–2307, 2010.

[33] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust object recognition with cortex-like mechanisms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 411–426, 2007.

[34] S. Frintrop, A. N¨uchter, H. Surmann, and J. Hertzberg, "Saliency-based object recognition in 3-D data," in *Proceedings of the International Conference on Intelligent Robots and Systems*, pp. 2167–2172, Sendai, Japan, October 2004.

[35] L. J. Croner and E. Kaplan, "Receptive fields of P and M ganglion cells across the primate retina," *Vision Research*, vol. 35, no. 1, pp. 7–24, 1995.

[36] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *Journal of the Optical Society of America A*, vol. 4, no. 12, pp. 2379–2394, 1987.

[37] T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio, *A Theory of Object Recognition: Computations and Circuits in the Feedforward Path of the Ventral Stream in Primate Visual Cortex*, Massachusetts Institute, Cambridge, MA, USA, 2005.

[38] C. Chang and C. Lin, "LIBSVM: a library for support vector machines," 2001, http://www.csie.ntu.edu.tw/%7Ecjlin/%20libsvm/.

[39] T. Liu, J. Sun, N. N. Zheng, X. Tang, and H. Y. Shum, "Learning to detect a salient object," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 596–603, Minneapolis, MN, USA, June 2007.

[40] Q. Yan, L. Xu, J. P. Shi, and J. Y. Jia, "Hierarchical saliency detection," in *Proceedings of the IEEEConference on Computer Vision and Pattern Recognition*, pp. 1155–1162, Portland, OR, USA, June 2013.

[41] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories," in *Proceedings of the IEEE Conference on Computer Vision*, pp. 178–178, Washington, DC, USA, 2004.

[42] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1597–1604, Miami, FL, USA, June 2009.

[43] M. M. Cheng, J. Warrell, W. Y. Lin, S. Zheng, V. Vineet, and N. Crook, "Efficient salientregion detection with soft image abstraction," in *Proceedings of the IEEE Conference on Computer Vision*, pp. 1529–1536, Sydney, NSW, Australia, December 2013.

[44] B. W. Jiang, L. H. Zhang, H. C. Lu, C. Yang, and M. H. Yang, "Saliency detection viaabsorbing Markov chain," in *Proceedings of the IEEE Conference on Computer Vision*, pp. 1665–1672, Sydney, NSW, Australia, December 2013.

[45] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2015.

[46] Z. Liu, W. Zou, and O. Le Meur, "Saliency tree: a novel saliency detection framework," *IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society*, vol. 23, no. 5, pp. 1937–1952, 2014.

[47] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, "Saliency filters: contrast based filtering for salient region detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 733–740, Providence, RI, USA, June 2012.

[48] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2814–2821, Columbus, OH, USA, June 2014.

[49] J. Kim, D. Han, Y.-W. Tai, and J. Kim, "Salient region detection via high-dimension alcolor transform," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 883–890, Columbus, OH, USA, June 2014.

[50] H. Z. Jiang, J. D. Wang, Z. J. Yuan, Y. Wu, N. N. Zheng, and S. P. Li, "Salient objectdetection: a discriminative regional feature integration approach," in *Proceedings of the IEEE-Conference on Computer Vision and Pattern Recognition*, pp. 2083–2090, Portland, OR, USA, June 2013.

[51] C. Yang, L. H. Zhang, H. C. Lu, X. Ruan, and M. H. Yang, "Saliency Detection via graph-based manifold ranking," in

Proceedings of the IEEE Conference on Computer Vision and Patternrecognition, pp. 3166–3173, Portland, OR, USA, June 2013.

[52] X. H. Li, H. C. Lu, L. H. Zhang, X. Ruan, and M. H. Yang, "Saliency detection via dense and sparse reconstruction," in *Proceedings of the IEEE Conference on Computer Vision*, pp. 2976–2983, Sydney, NSW, Australia, December 2013.

[53] S. Li, C. Zeng, S. Liu, and Y. Fu, "Merging fixation for saliency detection in a multilayer graph," *Neurocomputing*, vol. 230, pp. 173–183, 2017.

[54] A. Berg, *Shape Matching and Object Recognition*, Ph.D. Thesis, Department of Computer Science, University of California Berkeley, Berkeley, CA, USA, 2005.

[55] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision (ICCV)*, Kerkyra, Greece, September 1999.

[56] D. Khosla, D. J. Huber, and C. Kanan, "A neuromorphic system for visual object recognition," *Biologically Inspired Cognitive Architectures*, vol. 8, pp. 33–45, 2014.

[57] Y. F. Lu, T. K. Kang, H. Z. Zhang, and M. T. Lim, "Enhanced hierarchical model of object recognition based on a novel patch selection method in salient regions," *IET Computer Vision*, vol. 9, no. 5, pp. 663–672, 2015.

[58] M. Norizadeh, M. Shiri, and M. R. Daliri, "An enhanced HMAX model in combination with SIFT algorithm for object recognition," *Signal Image and Video Processing*, vol. 14, no. 3, 2020.

[59] T. Palmeri and I. Gauthier, "Visual object understanding," *Nature Reviews Neuroscience*, vol. 5, 2004.