Journal of Information Hiding and Multimedia Signal Processing ©2015 ISSN 2073-4212
Ubiquitous International Volume **6**, Number **6**, November **2015**

Visual Saliency Detection Based Object Recognition

Guanghua Gu*, Jiaojiao Zhu, Zexia Liu

School of Information Science and Engineering
YanShan University, Qinhuangdao, China
*Corresponding author
guguanghua@ysu.edu.cn

Yao Zhao

Institute of Information Science
Beijing Jiaotong University, Beijing, China

Received March, 2015; revised May, 2015

Abstract. *Visual object recognition is an open and challenging problem in computer vision. Note that fixing the position of the objects is one of the difficult problems in the object recognition task. Biological visual system tends to find naturally the most informative regions in a scene. In this paper, we present an object recognition approach based on the visual saliency. Firstly, the most salient region is obtained as the appropriate position of the object with the visual saliency detection method. Secondly, the dense SIFT (Scale Invariant Feature Transform) features are extracted from the detected salient region to form the image representation by using the LLC (Local-constrained Linear Coding) mapping scheme. Finally, the object recognition is completed with a linear SVM classifier. We evaluate the proposed method on the Graz-02, Caltech-256 and Pascal VOC 2006 datasets and verify the influence of the visual saliency.*
Keywords: Saliency detection; Object recognition; Local contrast; Convex hull; Image representation.

1. **Introduction.** Object recognition, i.e.determining the position and the class of an object within an image, has made impressive progress over the past few years. It is one of the fundamental challenges in computer vision. The objects present the large intra-class difference and inter- class similarity. The same object category can exhibit significant variations in appearance. For instance, the same dog can present very different poses. Besides, the pictures may also be affected by the confounding variables, such as the partial occlusions, deformations, viewpoints, object scales and the clutter backgrounds. Further more, inter-class similarity is also a major challenge in object recognition, such as the dog and the wolf, the dog and the cat. These factors make the object recognition much more difficult. Biological visual system tends to naturally find the most informative region in a scene. Object recognition is one of the most basic purposes of the visual system. The brain has zones that specialize in this task. Reliable saliency detection methods would be useful in many applications, such as unsupervised image segmentation [1], object recognition [2, 3], and motion detection [4]. In recent years, numerous computational saliency models have been proposed to simulate biological vision systems. These methods can be broadly classified into two major categories: bottom-up models and top-down models. Most existing bottom- up visual attention approaches considered the local contrast [1, 5-9]. It was developed to emphasize the uniqueness of a certain region by accentuating the

contrasts to neighbors. Besides, many models state that the central bias often exists in free scene viewing [10, 11]. The possible reason is that human often frame the object of interest near or in the center part of the image when they take a picture. The graph based visual saliency (GBVS) model [12] promoted higher saliency values in the center of the image, i.e. it has the center biases. The central bias is helpful when the salient object is in or near the center of the image. However, it cannot well detect the visual saliency when the object is off the center of the image. Compared to the center prior, Xie et al [13] first proposed the convex hull center prior. It can improve the robustness of the object location. In this paper, the convex hull center prior is used instead of the image prior. Some methods can highlight the salient object boundaries, but fail to uniformly map the entire salient region[5, 12]. Liu et al [6] proposed a set of novel features, multi-scale contrast, center-surround histogram and color spatial distribution, to describe a salient object, and then a conditional random field was learned to combine these features for salient object detection. Goferman et al [14] presented a context-aware saliency detection algorithm based on four psychological principles: local low-level considerations, global considerations, visual organization rules and high-level factors. The smoothness prior is used to modify the initial saliency map. The bag-of-features (BOF) framework has been demonstrated to be one of the most successful approaches to scene categorization and object recognition. The BOF strategy is simple and effective. However, since the classical BOF representation is applied to the entire image, which gives the chance for background clutter to disturb, or even overwhelm the object information, many content-irrelevant local features may result in the noisy and non- descriptive visual words in images. Moreover, it also discards the spatial information of local features in the images. To overcome it, the spatial pyramid matching (SPM)[15] method was proposed. It used a sequence of grids at different scales to partition the image into sub-regions, and then computed a BOF histogram for each sub-region. The representation of the whole image is the concatenation vector of all the histograms. However, most of the methods are only efficient on the pictures with the simple backgrounds and big obvious objects. Besides, the state-of-art object recognition method is based on exhaustive search over the image to find the best object position. However, as the total number of images and windows to evaluate in an exhaustive search is huge and growing, it is necessary to constrain the computation per location and the number of locations. Sande et al [16] adapted segmentation as a selective search, which greatly reduces the number of locations to be considered. Lu et al [17] proposed a superpixel-level object recognition in the BOF framework. It introduced local learning classifiers that turned the highly non-linear classification problems into the multiple local linear problems. In this paper, the human visual perception mechanism is applied to the object recognition framework. A bottom-up attention-based saliency model is used to identify the regions that are likely to contain objects. The convex hull center[13] instead of the image center is used to calculate the saliency maps, and the OSTU [18] segmentation algorithm is used to get the most salient regions. Then, features are extracted from this salient region, and a model is trained for object classification.

2. **Implementation.** As well known, one of the difficult problems in object recognition is the fixation of the object. If the position of the object can be located correctly, the recognition will become easier. In this paper, the human visual perception mechanism is applied into the object recognition framework. It includes two steps: saliency detection and object classification. First, we utilize a bottom-up saliency detection method to get the most salient regions of the image. Then, we extract the dense SIFT descriptors of the salient region, and a linear SVM was trained for object classification.

2.1. **Saliency computation.** The task of saliency detection is to identify the most important and informative part of an image. Given a color image I, we first over-segment it into a number of superpixels using simple Linear Iterative Clustering (SLIC) method [19]. For each superpixel, we apply the local contrast prior [20] and the convex-hull based center prior [13] to get the initial saliency map. Then, the smoothness prior [8] is used to refine the initial map and get the final saliency map. At last, we utilize the adaptive threshold to get the most salient regions.

2.1.1. *Local contrast.* Many methods [6, 12, 21] take account of the local contrast. They computed the regional contrast based on the feature dissimilarity and distance coherence. In this paper, we propose a local contrast according to the regional consistency [20]. If the regions are more similar in feature and closer in space, these regions will have close saliency values. For a superpixel i, its contrast saliency compared to all other superpixels is defined as:

$$A(i) = \sum_{j=1}^{N} D_F(i,j) \times D_S(i,j) \times f(j) \tag{1}$$

The first term $D_F(i,j)$ indicates the feature dissimilarity of two superpixels i and j. It is simply defined as the difference between the feature values of the two superpixels:

$$D_F(i,j) = |F(i) - F(j)| \tag{2}$$

In Eq.(1), the second term $D_S(i,j)$ indicates the spatial relationship of the two superpixels.

$$D_S(i,j) = \exp\left[\frac{\left(\frac{x_i - x_j}{M}\right)^2 + \left(\frac{y_i - y_j}{N}\right)^2}{2\sigma_0^2}\right] \tag{3}$$

Where (x_i, y_i) is the centroid of the superpixel i defined as the position of the superpixel. $M \times N$ is the size of the image I. σ_0 is the scale parameter which controls the power of the spatial weight.
In Eq.(1), the last term $f(j)$ represents the relative size of the compared superpixel j. The bigger regions have greater impact on the superpixel i. It denotes the size of the region. The relative size of a superpixel j is defined as.

$$f(j) = \frac{S(j)}{S(I)} \tag{4}$$

As shown in the Eq.(1), the local contrast saliency of superpixel i is determined by dissimilarity, closeness and relative size of the compared region. That is, the bigger, the nearer and more dissimilar regions will have greater influence on the current superpixel.

2.1.2. *Central bias.* Some works [10, 12, 22] attempted to take a central bias for saliency detection. It is helpful when the salient object is in or near the center of the image. However, these methods cannot well detect the visual object when it is far from the center of the image. We use the convex-hull-based center proposed by Xie *et al.,* [13] instead of the image center to get the center prior map. Some salient points are detected by the color boosted Harris point operator [23]. Then we compute a convex hull enclosing the interest points to estimate the location of the salient region. We apply the centroid of the convex hull to get the convex-hull-based center prior map. Given the center (x_0, y_0), the saliency of superpixel i is defined as:

$$B(i) = \exp\left[-\frac{\|x_i - x_0\|^2}{2\sigma_x^2} - \frac{\|y_i - y_0\|^2}{2\sigma_y^2}\right] \tag{5}$$

Where x_i and y_i denote the mean horizontal and vertical positions of the superpixel i, respectively. σ_x and σ_y indicate the horizontal and vertical variances. Here we set $\sigma_x = \sigma_y$. Inspired by the structure adapted from the Feature Integration Theory [34], we fuse the Eq.(1) and Eq.(5) to get the initial saliency map.

$$S_0(i) = A(i) \times B(i) \tag{6}$$

2.1.3. *Final saliency map.* To refine the saliency map, we introduce a smoothness prior [8] to take the interaction between image elements into account. It encourages the neighboring pixels taking the same labels. Here a saliency cost function is defined to encode the smoothness prior:

$$E(S) = \sum_i \left(S(i) - S_0(i)\right)^2 + \lambda \sum_{i,j} \omega_{ij} (S(i) - S(j))^2 \tag{7}$$

Where $S(i)$ and $S(j)$ correspond to the saliency values of superpixel i and j respectively, $S_0(i)$ is the initial saliency value for superpixel i. $\omega_{ij} = \exp(-\frac{\|c_i - c_j\|}{2\sigma_\omega^2}) \in W$ is the weight between two linked superpixels i, j. c_i indicates the mean of the superpixel i in the CIELAB color space. In Eq.(7), the first term $\sum_i \left(S(i) - S_0(i)\right)^2$ is the fitting constraint, which means a good saliency map should be close to the initial saliency map. The second term $\sum_{ij} \omega_{ij} (S(i) - S(j))^2$ is the smoothness constraint, which means that the neighboring superpixels should have similar saliency values. The optimal saliency values of superpixels are computed by minimizing the Eq.(7), which can be solved by setting the derivative of the above function with respect to S to be zero.
The resulting solution is:

$$S = \mu(D - W + \mu I)^{-1} S_0 \tag{8}$$

Where D is the diagonal degree matrix with $d_{ij} = \sum_j \omega_{ij}$ and $\mu = \frac{1}{2\lambda}$. The saliency value is normalized to [0, 1]. Then the OSTU[9] algorithm is used to get the most salient regions.

2.2. **Object classification.** In recent years, the bag-of-words based framework has been demonstrated to be one of the most successful approaches to the scene and object recognition. The BOF approach discards the spatial order of local descriptors, which severely limits the descriptive power of the image representation. To overcome it, the extension of the BOF model, addressed as the spatial pyramid matching (SPM), has made a remarkable success on a range of the image classification. The SPM method partitions an image into $2^l \times 2^l$ sub-regions and computes the BOF histograms of local features for each sub-region. After getting the salient regions of the image, we use the SPM framework on the salient regions to classify the objects.

2.2.1. *Feature extraction.* Most recent approaches adopt local invariant features as an effective image representation, because they can well describe and match instances of objects or scenes under a wide variety of image scaling, rotations, illuminations and viewpoints. The SIFT descriptor [25] is one of the most widely used descriptors due to its good performances. Given a feature point, the SIFT descriptor computes the gradient vector for each pixel in the feature points neighborhood and builds a normalized histogram of gradient directions. To achieve better discriminative ability, we extract the dense SIFT descriptors for the salient regions.

2.2.2. *Feature coding.* The classification result is heavily influenced by the coding methods. Vector quantization (VQ) is widely used to encode the feature descriptors on the condition of the least square fitting:

$$\arg\min_{V} \sum_{i=1}^{N} \|x_i - Vu_i\|^2 \tag{9}$$

$$s.t. \ \left\|\underset{i}{u}\right\|_{l_0} = 1, \ \left\|\underset{i}{u}\right\|_{l_1} = 1, \ \underset{i}{u} \geq 0, \ \forall i$$

where $X = [x_1, x_2, ..., x_N] \in R^{D\times N}$ is a set of D-dimensional local descriptors extracted from an image. is a codebook with M entries. $U = [u_1, u_2, ..., u_N]$ represents the set of codes for X. The constraint condition $\|u_i\|_{l_0} = 1, \ \|u_i\|_{l_1} = 1, \ u_i \geq 0$ restricts that there is only one non-zero element in each code u_i. In VQ, each descriptor is represented by a single basis in the codebook. Due to the large quantization errors, the VQ code for similar descriptors might be very different. Besides, the VQ process ignores the relationships between the different bases. Wang *et al* [26] proposed a locality-constrained linear coding (LLC) instead of VQ coding in traditional SPM. LLC utilizes the locality constraints to project each descriptor into its local-coordinate system, and the projected coordinates are integrated by max pooling to generate the final representation.

$$\min_{V} \sum_{i=1}^{N} \left\|\underset{i}{x} - V\underset{i}{u}\right\|^2 + \lambda \left\|\underset{i}{d} \odot \underset{i}{u}\right\|^2 \tag{10}$$

$$s.t. \ 1^T \underset{i}{u} = 1, \forall i$$

where $\odot$ denotes the element-wise multiplication; $d_i = \frac{dist(x_i,V)}{\sigma} \in R^M$ is the locality adaptor that gives different freedom for each basis vector proportional to its similarity to the input descriptor x_i; $dist(x_i, V) = [dist(x_i, v_1), , dist(x_i, v_M)]^T$, where $dist(x_i, v_j)$ is the Euclidean distance between x_i and v_j; is used to adjust the weight decay speed for the locality adaptor. In the LLC, each descriptor is more accurately represented by multiple bases, and LLC code captures the correlations between similar descriptors by sharing bases. For object categorization, SVM is the preferred classifier as a powerful technology in machine learning. In this paper, we use a linear SVM to complete the training and testing.

3. **Experiments and results.**

3.1. **Datasets and setup.** We evaluate the proposed method on three widely used datasets. The first one is the Graz-02 dataset. It contains three object categories, i.e. bike, car and person. Each category has 300 images of size 640×480 pixels and corresponding pixel-level foreground/background annotations. In the experiment, we resize the images into 320×240 pixels. The average size ratio of the object versus the image is: 0.22 for bikes, 0.177 for persons, and 0.09 for cars [27]. The second one is the Caltech-256 dataset, an extension of the Caltech-101 dataset. It consists of 256 object categories, each of that has at least 80 images. The total image number of Caltech-256 dataset is 30686. Compared to Caltech-101, Caltech-256 possesses larger inter-class and intra-class variability, and thus it is more challenging. The last one is the Pascal VOC 2006 dataset. It includes 10 categories (bicycle, bus, car, cat, cow, dog, horse, motorbike, person, and sheep). In the trainval set, there are 2618 images, containing 4754 annotated objects. The image of the Caltech-256 dataset contains only one object category with simple background. The

TABLE 1. Classification results on the Graz-02 datasets

	Graz1 (%)	Graz2 (%)	Graz-02 (%)
Without saliency	70.84	94.75	82.56
With saliency	76.21	91.02	84.33

image background of the Graz-02 dataset is more clutter than the Caltech-256. However, in the PASCAL VOC dataset, images are always with multiple objects in very clutter backgrounds. In the saliency detection process, we set the number of superpixel nodes N=200 in all the experiments. In our implementation, we empirically chose $\sigma_0^2 = 0.2$ in Eq.(3),$\sigma_x^2 = \sigma_y^2 = 0.15$ in Eq.(5)and $\lambda = 25$ in Eq.(7). In the object recognition process, we extract the dense SIFT descriptors on the salient region. The SIFT descriptor of 16× 16 pixel patches are computed over a grid with spacing of 8 pixels. After the feature extraction, the unsupervised K-means clustering method is performed to build the vocabulary of the visual words. 500 is chosen as the dictionary size M in our implementation. The descriptors are coded by LLC coding to the nearest 5 codebook centers. Max pooling is used in the SPM with *l=0, 1, 2*. A linear SVM is trained for classification. In addition, we compare the performance of the proposed method with the classical SPM framework without saliency detection. Without special instructions, we randomly select 50% images from each category as the training set and the remaining 50% as the testing set. All the final results are reported as the average of 10 individual runs.

3.2. **Results on Graz-02 dataset.** As we all know, if the object occupies most part of the image, it is not necessary to do the saliency detection. Our methods are more efficient to the images of smaller objects. In order to verify our method, we divide the Graz-02 dataset into two subsets, named as Graz1 and Graz2. Graz1 consists of 150 images with the relative small objects (such as the first three rows of Fig.1), and Graz2 is composed of the rest 150 images with the relative big objects (the last three rows of Fig. 1). For the Graz1 and Graz 2, we randomly select 75 images from each category for training and the rest 75 images for testing. For the whole Graz-02 dataset, we use the setting as [27] in our experiments. We take 150 odd-numbered images of each category as the training set, and the remaining 150 even-numbered images as the testing set. We compare the saliency map of Ittis [5], Yangs [8] and our method. Fig. 1 gives some examples of the saliency detection results.

The saliency detection methods can get the most salient regions of the image and reduce the influence of the clutter background. We evaluate the performance of our method on Graz 1, Graz 2 and the whole Graz-02 datasets. The average classification results are displayed in Table 1.

3.3. **Extraction Technique.** As can be seen from the Table 1, saliency detection has a great influence on the object classification. On Graz 1, the images with small objects are difficult to detect. It contains clutter backgrounds and the objects are not obvious. The proposed visual saliency based method can locate on the salient region of the image and improve the classification result. With saliency, the classification result can achieve 76.21%, while without saliency the result is only 70.84%. However, on the Graz 2 dataset, the objects are relatively bigger. With saliency detection the average classification accuracy is a little lower than that without saliency detection. The saliency detection may not locate on the object properly. As shown in the last column of Table 1, on the whole Graz-02 dataset, with saliency detection the improvement is 1.77%. It is slightly better than that without saliency. We also evaluate the influence of different saliency detection methods on the Graz-02 dataset. As shown in Table 2, we compare the results of Ittis

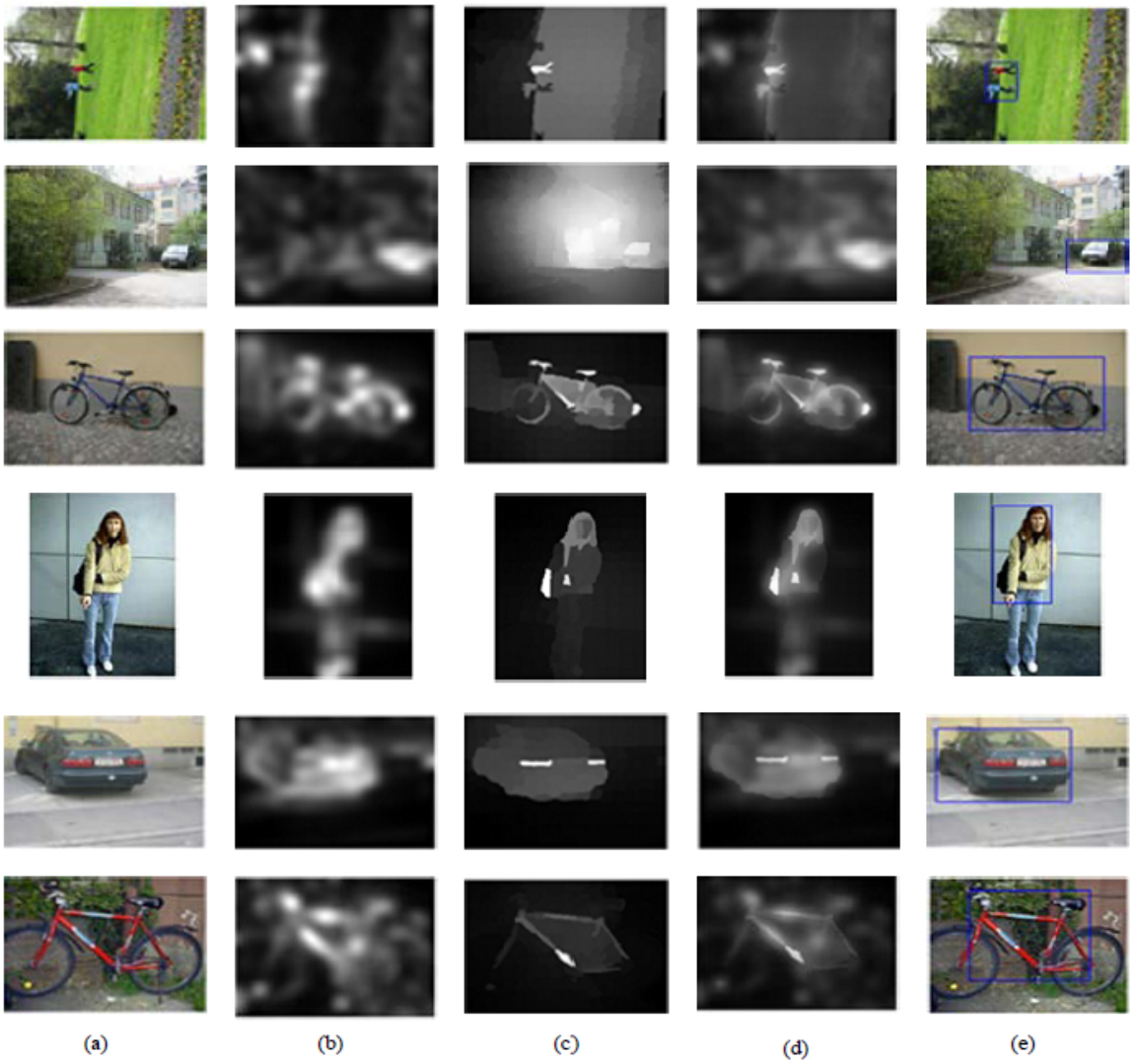

FIGURE 1. Saliency detection results on the Graz-02 datasets. First three rows are some examples of the Graz1 with relatively small objects and the last three rows are examples from Graz2. (a) Original images; (b) Ittis saliency maps; (c) Yangs saliency maps; (d) our saliency maps; and (e) are the salient regions detected by our methods.

TABLE 2. Classification accuracy comparison on Graz-02 dataset

	Car (%)	Bike (%)	Person (%)	Graz-02 (%)
Without saliency	86.7	80.7	80.3	82.56
Ittis [5] saliency	85.8	79.0	76.8	79.85
Yangs [8] saliency	86.3	81.0	80.0	82.43
Our saliency	88.0	85.3	79.7	84.33

[5] saliency, Yangs [8] saliency, our saliency and the results of without saliency detection. Table 2 gives the classification results on the car, bike and person categories and the average results of the whole Graz-02 dataset respectively. From Table 2 we can see, our approach achieves the best performance. For the car and bike categories, our method gains the highest classification rate, Yangs [8] method is appropriate to that of without saliency detection. However, for the person category, the result of our method is similar

to Yangs[8] saliency and without saliency. The result of Ittis [5] saliency is lower than the others. But on the whole dataset, our methods can achieve the highest rate of 84.33%. Our saliency detection result outperforms Ittis [5] saliency by 3.5%. Compared to Yangs [8] saliency and without saliency, our result improves nearly 2%. In general, our approach is the best.

3.4. **Results on the subset of Caltech-256.** We further evaluate the proposed method on the Caltech-256 dataset. We select ten categories from the Caltech 256 as a subset, containing aeroplanes, motorbikes, cars, binoculars, faces, computer-mouse, dogs, coins, leaves and tomatoes. For each category, we randomly select 80 images in our experiments. Among them, 56 images are for training and the rest 24 images are for testing. In the experiments, we analyze the following two aspects: one is the classification performance with the number of the training images; the other is the relationship between the classification performance and the saliency detection.

In the above dataset, most images are with a single object in a simple background. Some

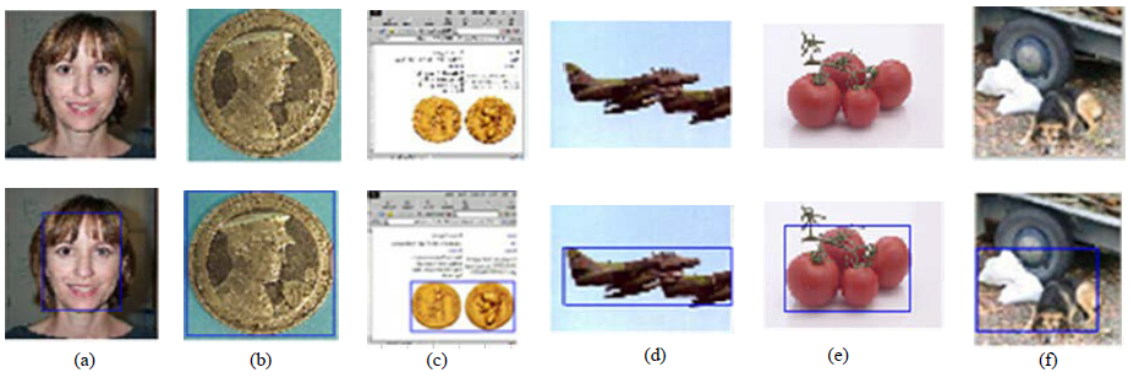

FIGURE 2. Some examples of the Caltech-256 dataset: (a) the face category; (b) and (c) the coin category; (d) the aeroplane category; (e) the tomato category; (f) the dog category.

examples and their saliency detection results (the blue bounding box) are shown in Fig. 2. Most objects occupy the main parts of the images, such as the face in Fig. 2 (a) and the coin in Fig. 2 (b). In addition, some of them are with multiple instances of the same object category, like Fig. 2 (c)-(e). But in some instance, the background is cluttered and the object is not salient to recognize, for example the dog in Fig. 2 (f). The white thing is more salient than the dog. It may influence the recognition result.

For the subset of Caltech 256 dataset, the overall result is very high. Fig. 3 gives the confusion matrix of the average experiment results. The average classification result of the ten categories is 86.25% with saliency (Fig. 3 (a)) and 86.33 % without saliency (Fig. 3 (b)). The two results are very close for the dataset. It can be seen, in some categories (like aeroplanes, motorbikes, cars, faces and leaves), the classification rates are all above 93%. Some of them are nearly 100%. The result of the computer-mouse category with saliency is lower than that without saliency. Some of the computer-mice are similar to tomatoes in shape or appearance. It produces much confusion with the binocular and the tomato categories. But with saliency, the improvement is obvious in the coin and the tomato categories. For the coin category, the one with saliency outperforms the other with no saliency by 5%; especially for the tomato category, the improvement can be as high as 15%. With saliency detection, the confusion with the computer-mouse category is greatly reduced. The two results of the dog category with and without saliency are

relatively lower than the other categories. The main reason is that the dog has diverse appearance and deformation. It is hard to recognize. Even if the average results of the ten categories are very close, the advantage of saliency detection is highly obvious for some categories (like coin and tomato). To evaluate the classification results with differ-

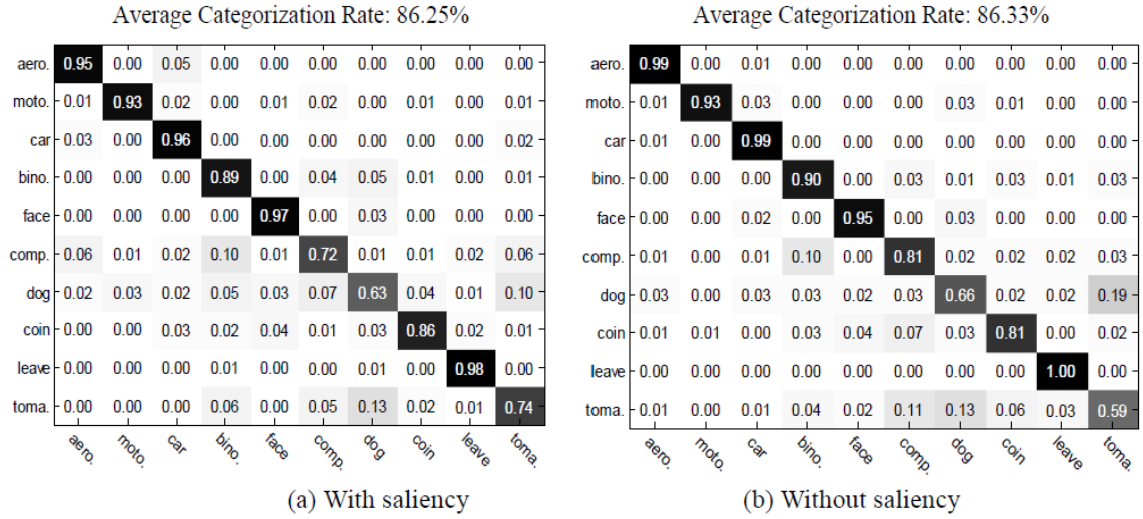

FIGURE 3. Confusion matrixes of the ten categories from Caltech 256.

ent training image numbers. We randomly select 8, 16, 24, 32, 40, 48, 56, 64 images from each categories respectively for training and the rest images for testing. As is shown in Fig. 4, the performance increases gradually when the training number grows from 8 to 64 in each category. When the train test ratio below 0.3, the result of applying saliency detection is a little lower than that without saliency. With the increase of training images, the method applying saliency detection is superior to ones without saliency. The more there are training images, the more obvious is the saliencys superiority. The classification accuracy changes from 73.61% to 88.75% with the ratio changing from 0.1 to 0.8, while without saliency, the result varies from 75.45% to 87.71%. As we all know, the more training samples there are, the more close the models are to the real objects. The classification results will be higher.

3.5. **Results on the PASCAL VOC2006 dataset.** The standard PASCAL VOC 2006 dataset includes 10 categories (bike, bus, car, cat, cow, dog, horse, motorbike, person, and sheep). This dataset is extremely challenging because all the images are daily photos obtained from Flicker where the size, viewing angle, illumination, appearances of objects and their poses vary significantly, with frequent occlusions and a wide range of appearance, shapes and scales. Most of the images contain multiple objects on cluttered backgrounds. Note that the dataset is not suitable for our salient object recognition, because some objects are not salient (too small or partially occluded). In the experiments, we select six categories (i.e. bike, car, cat, cow, motorbike and sheep) from the trainval set to make a new subset. In our implementation, 200 images are chosen randomly from each category and divided into two separate image sets, i.e. 100 for training and 100 for testing respectively. Fig. 5 gives the confusion matrix of the average experiment results.

Note that with saliency (Fig. 5 (a)) the average classification accuracy is 64.67% and 4.16% lower than that without saliency (Fig. 5 (b)). On this dataset, the saliency detection is bad to the final result. The main reason is that for the images with multiple

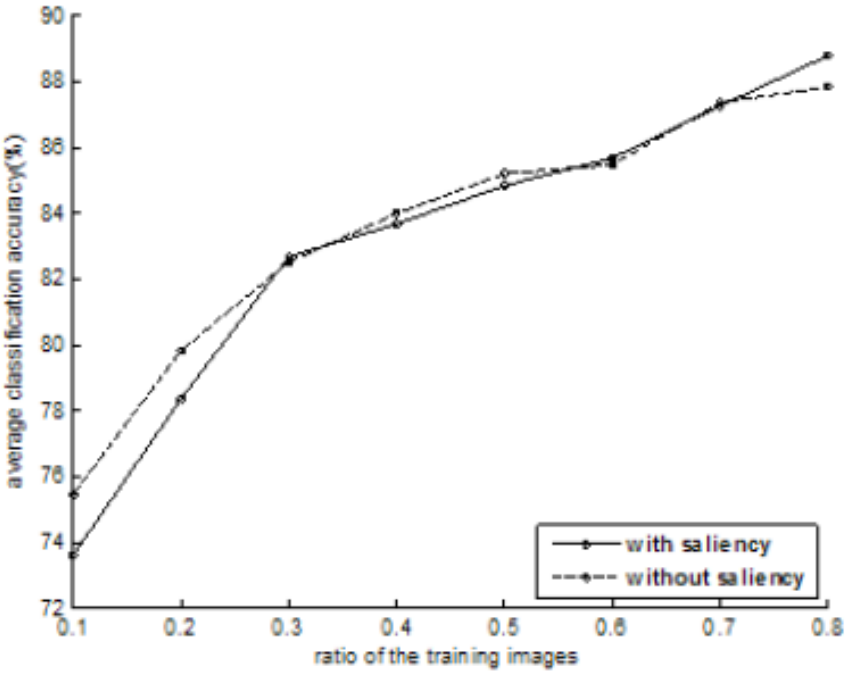

FIGURE 4. Performance with different training images.

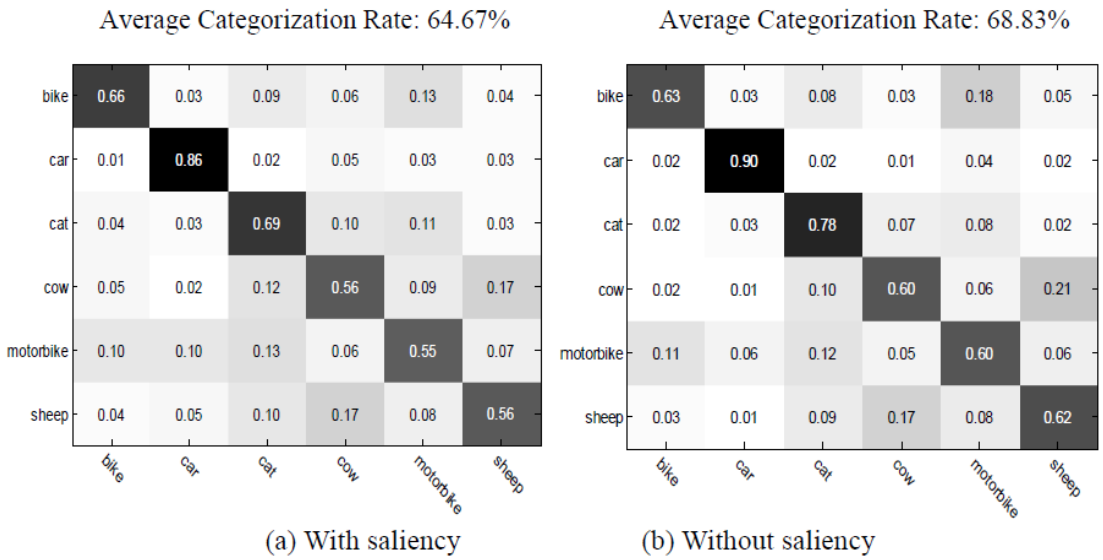

FIGURE 5. Confusion matrix of Pascal VOC 2006 dataset

objects in the very complex background, the bottom-up saliency cannot locate on the object correctly. Clearly, it cannot recognize the object.

Some examples of this dataset are shown in Fig.6. For some images, the saliency detection can not locate on the object of interest correctly. Most images contain multiple object categories, such as Fig.6 (a)-(b). It is hard to locate on the object of interest without prior knowledge. In Fig.6 (c), the white cat is very similar to the background. And Fig.6 (d) is a cat in a disordered background. Recognizing the cat is very difficult. Besides, even the locations are precisely detected, some objects are also very hard to recognize. Fig.6 (e) and (f) are two similar cases, i.e. a person riding a bike and riding a motorbike

respectively. It is hard to tell what he was riding. Fig.6 (g) is a cow, which has a great similarity with the sheep. Fig.6 (h) is a sheep under the setting sun.

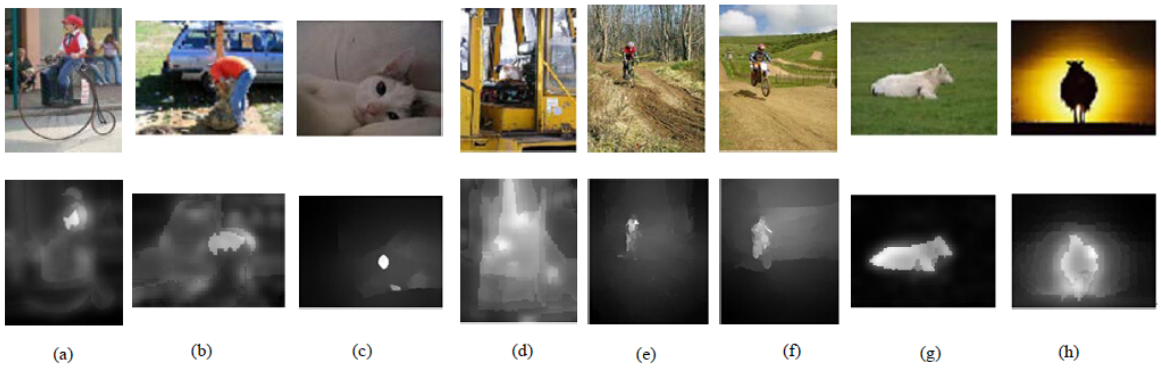

FIGURE 6. Figure 6 Some failure examples of the VOC dataset of our methods: (a) and (b) a picture with multiple objects; (c) a cat in a similar background; (d) a cat in a clutter background; (e) a person rides a bike; (f) a person rides a motorbike; (g) a cow; (h) a sheep.

4. **Discussion.** In section 4, we evaluate our method on three widely used datasets. It demonstrates that the performance of the saliency detection has great influence on the final classification accuracy. Knowing the positions of objects will be convenient to the recognition. However, if the object's position is not correctly located, the recognition will be hard. In this section, we will discuss the different results of our method on these datasets as the following four aspects.

(1) It is sure that reliable saliency detection methods would be useful for object recognition. The proposed visual saliency can locate on the salient regions of the image and improve the classification results. As shown in the first row of Fig. 1, the image contains much unrelated information, the two people are very small to recognize. The saliency detection method can locate on the salient object regions and reduce the interference of backgrounds. In Graz 1, the images with clutter backgrounds and relatively small objects are difficult to recognize. The visual saliency can reduce the influence of the clutter background and improve the object recognition results. For example in Table 1, the classification rate with saliency can improve 5.87% compared to non-saliency on Graz1. Different saliency detection methods have different impact on the classification results. If the salient region is detected properly, the recognition will be easy. On the contrary, if the position is located wrongly, it is impossible to recognize the object. The Itti's [5] saliency model is simple and easy to complete. However, it usually responds to numerous unrelated backgrounds, and has high saliency values on the boundary of the objects, yet cannot uniformly highlight the objects (for example in Fig.1 (b)). The classification results are lower than the others. For the local contrast prior, Yang's [8] saliency was defined as the spatial weighted contrast. It may wrongly highlight part of the background, such as the second row of Fig.1 (c). In our method, we refine the local contrast factor according to the regional consistency. That is, if the regions are more similar in feature and closer in space, these regions will have the close saliency values. Our method can uniformly highlight the whole object region and simultaneously suppress the background region effectively. On the whole, our method can achieve the best performance as shown in Table 2.

(2) For the image with big objects, the saliency detection becomes invalid. In Graz 2

and Caltech 256 datasets, most of the objects are occupying the main parts of the image. Some of them even fill up the whole image (like the Fig.2 (a) and (b)). In this case, the saliency detection is not necessary and it may not locate on the object properly. Some of the objects are detected with only a small part. For example, in the fourth row of Fig. 3, the saliency detection only locates on a part of the person. And in the last row of Fig. 1, the bike is also partly detected. Some object information is missing. In this case, the classification results must be influenced. However, with the whole image, it is easy to recognize the object. As demonstrated on Graz 2 dataset of Table1, the classification result with saliency is a little lower than that without saliency. In Fig.3, the average results of with and without saliency detection are very close. The advantage of saliency detection is not obvious. In order to enhance the classification rates, it is necessary to improve the saliency algorithm and get more accurate object region.

(3) Due to lack of prior knowledge, the bottom-up visual saliency cannot locate on the objects of interest correctly. If the location is wrongly detected, it is impossible to recognize the object. On the Pascal VOC dataset, many images contain multiple objects. In Fig.6 (a), the person is more salient compared to the bike. If we want to recognize the bike, it may be wrong. Fig. 6 (b) contains a person and a car. It also cannot locate on the person or the car region precisely. The final classification result is heavily influenced by the regions detected by the visual saliency. Sometimes, the saliency detection result includes much useless information and it may also influence the classification result. In Fig. 2 (f), the white thing is more salient than the dog related to the whole image. The process of saliency detection may be located in the uninformative background rather than the object of interest. In Fig.6 (c), the cat is too similar to the background. The eye is more salient related to the whole image. In this case, the saliency detection locates on he eye region. Only with the eye part, it is difficult for us to identify it as a cat or other animals. However, with the whole image, it is easy to recognize the cat. Beyond that, the background clutter may still affect the detection results. In Fig.6 (d), the cat is in a very clutter background and the saliency detection methods cannot locate on the cat region. In these cases, the saliency detection will be bad to the result and our method may not recognize the object correctly. Some prior information must be included to deal with these problems.

(4) Besides, the inter-class similarity may also influence the classification rate. For example in Fig.6 (e) a person rides a bike, and in Fig.6 (f) a person rides a motorbike. At first glance, the two instances are very similar. Even in some situations, the person is more salient than the bike or motorbike. The saliency detection method cannot locate on the bike or motorbike region correctly. In addition, some object categories are very similar in visual appearance, like the dog and the cat or the cow and the sheep categories. We can not correctly recognize the salient object as a cow or a sheep or some other animals in Fig.6 (g). And in Fig.6 (h), because of the illustration and perspective, it is hard to recognize the salient object. So even if the positions are rightly detected, the objects may also be identified wrongly. Adding in some prior knowledge and exploiting more powerful visual features should be helpful to improve the object recognition accuracy.

5. **Conclusions.** In this paper, we presented a visual saliency-based object recognition method. The objects are always conspicuous in the image, so we use a bottom-up visual saliency method to get the most salient regions of the images. Then we perform the object recognition on this salient region. It can greatly reduce the unrelated information on the background and improve the classification result. However, when the background is highly clutter, due to lack of top-down prior knowledge, the bottom-up saliency algorithms usually respond to numerous unrelated low-level visual stimuli and thus missing the object

of interest. It is necessary to improve the saliency algorithm and get more accurate object region in the future. Besides, in order to deal with the multiple object recognition in complex scenes, some prior information must be included.

Acknowledgment. This work was partly supported by Natural Science Foundation of China (No. 61303128), Natural Science Foundation of Hebei Province (No. F2013203220, No. F2014203132), Key Foundation of Hebei Educational Committee (ZD2015095), Youth Foundation of Hebei Educational Committee (Q2012047), and Program for Plan of Science and Technology of Qinhuangdao City (201401A017).

REFERENCES

[1] E. Rahtu, J. Kannala, M. Salo, and J. Heikkila, Segmenting Salient Objects from Images and Videos, I*n Proc. of European Conference on Computer Vision*, vol. 6315, pp. 366-379, 2010.

[2] U. Rutishauser, Christof Koch, and Pietro Perona, Is bottom-up attention useful for object recognition?, *In Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 37-44, 2004.

[3] N. Z. L. Yang, M. Chen,Y. Yang, J. Yang, Categorization of Multiple Objects in a Scene Using a Biased Sampling Strategy, *International Journal of Computer Vision*, pp. 1-18, 2013.

[4] C. Liu, P. C. Yuen, and G. P. Qiu, Object motion detection using information theoretic spatio-temporal saliency, *Pattern Recognition*, vol. 42, pp. 2897-2906, 2009.

[5] C. K. L. Itti, E. Niebur, A model of saliency based visual attention for rapid scene analysis, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pp. 1254-1259, 1998.

[6] T. Liu, Z. J. Yuan, J. A. Sun, et al., Learning to Detect a Salient Object, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 33, pp. 353-367, 2011.

[7] R. H. Achanta, S., Estrada, F., Susstrunk, S., Frequency-tuned salient region detection, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1597-1604, 2009.

[8] C. Yang, L. H. Zhang, and H. C. Lu, Graph-Regularized Saliency Detection With Convex-Hull-Based Center Prior, IEEE Signal Processing Letters, vol. 20, 2013.

[9] F. Y. Tian H., Zhao Y., Lin W., Ni R., Zhu Z., Salient Region Detection by Fusing Bottom-Up and Top-Down Features Extracted From a Single Image, *IEEE Trans. on Image Processing*, vol. 23, pp. 4389-4398, 2014.

[10] T. Judd, K. Ehinger, F. Durand, and A. Torralba, Learning to Predict Where Humans Look, *In Proc. of IEEE International Conference on Computer Vision*, pp. 2106-2113, 2009.

[11] M. Bindemann, Scene and screen center bias early eye movements in scene viewing, Vision Research, vol. 50, pp. 2577-2587, 2010.

[12] J. Harel, C. Koch, and P. Perona, Graph-based visual saliency, Advances in Neural Information Processing Systems, vol. 19, pp. 545-552, 2007.

[13] Y. L. Xie, H. C. Lu, and M. H. Yang, Bayesian Saliency via Low and Mid Level Cues, *IEEE Trans. on Image Processing*, vol. 22, pp. 1689-1698, 2013.

[14] S. Goferman, L. Zelnik-Manor, and A. Tal, Context-aware saliency detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 1915-1926, 2012.

[15] S. Lazebnik, C. Schmid, and J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, *In Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 2169-2178.

[16] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, Selective Search for Object Recognition, *International Journal of Computer Vision*, vol. 104, pp. 154-171, 2013.

[17] H. C. Lu, X. J. Feng, X. H. Li, and L. H. Zhang, Superpixel level object recognition under local learning framework, *Neurocomputing*, vol. 120, pp. 203-213, 2013.

[18] S. H. Jiao, X. G. Li, and X. Lu, An improved Ostu method for image segmentation, *The 8th International Conference on Signal Processing*, vol. s 1-4, pp. 966-969, 2006.

[19] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, SLIC Superpixels Compared to State-of-the-Art Superpixel Methods, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 2274-2281, 2012.

[20] J. Zhang, J. Ding, and J. Yang, Exploiting global rarity, local contrast and central bias for salient region learning, *Neurocomputing*, 2014.

[21] M. M. Cheng, G. X. Zhang, N. J. Mitra, X. L. Huang, and S. M. Hu, Global Contrast based Salient Region Detection, *In Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 409-416, 2011.

[22] Q. Zhao and C. Koch, Learning a saliency map using fixated locations in natural scenes, *Journal of Vision*, vol. 11, 2011.

[23] J. van de Weijer, T. Gevers, and A. D. Bagdanov, Boosting color saliency in image feature detection, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 150-156, 2006.

[24] A. M. Treisman and G. Gelade, A feature-integration theory of attention, *Cognitive psychology*, vol. 12, pp. 97-136, 1980.

[25] D. G. Lowe, Object recognition from local scale-invariant features, *in In Proc. of IEEE International Conference on Computer Vision*, 1999, pp. 1150-1157.

[26] J. J. Wang, J. C. Yang, K. Yu, F. J. Lv, T. Huang, and Y. H. Gong, Locality-constrained Linear Coding for Image Classification, *In Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3360-3367, 2010.

[27] J. Yang, N. N. Zheng, J. Yang, M. Chen, and H. Chen, A Biased Sampling Strategy for Object Categorization, *In Proc. of IEEE International Conference on Computer Vision*, pp. 1141-1148, 2009.