IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 7, 2019, accepted March 19, 2019, date of publication April 1, 2019, date of current version April 16, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2908448

Scene Categorization Model Using Deep Visually Sensitive Features

JING SHI[1], HONG ZHU[1], SHUNYUAN YU[2], WENHUAN WU[1,3], AND HUA SHI[4]

[1]Institute of Automation and Information Engineering, Xi'an University of Technology, Xi'an 710048, China
[2]Institute of Electronic and Information Engineering, Ankang University, Ankang 710025, China
[3]Institute of Electrical and Information Engineering, Hubei University of Automotive Technology, Shiyan 442002, China
[4]Institute of Sciences, Xi'an Technological University, Xi'an 710021, China

Corresponding author: Hong Zhu (zhuhong@xaut.edu.cn)

This work was supported in part by the Nature Science Foundation of China under Grant 61801005, Grant 61771386, and Grant 61673318.

ABSTRACT Visually sensitive regions in the scene are thought to be important for scene categorization. In this paper, we propose to utilize the important visually sensitive information represented by deep features for scene categorization. Specifically, the context relationship between the objects and the surroundings is fully utilized as the main basis for judging the content of the scene, and combining with the deep convolution neural networks (CNNs), a scene categorization model based on deep visually sensitive features is constructed. First, the saliency regions of the scene images are marked according to the context-based saliency detection algorithm. Then, the original images and the corresponding visually sensitive region detection images are superimposed to obtain the visually sensitive region enhancement images. Second, the deep convolution features of the original images, the visually sensitive region detection images, and the visually sensitive region enhancement images are extracted through the deep CNNs pre-trained on the large-scale scene dataset Places. Finally, considering that the deep features extracted by different layers of the convolution network have different capabilities of discrimination, the fusion features are generated from multiple convolution layers to construct visually sensitive CNN model (VS-CNN). In order to verify the effectiveness of the proposed model, the experiments are conducted on the five standard scene datasets, i.e., LabelMe, UIUC-Sports, Scene-15, MIT67, and SUN. The experimental results show that the proposed model is effective and has good adaptability. Especially, our categorization performance is superior to many state-of-the-art methods for a complex indoor scene.

INDEX TERMS Scene categorization, deep convolution networks, visually sensitive features, categorization model.

I. INTRODUCTION

Scene categorization means to classify scene images into different semantic classes based on their contents. As one of the main problems in computer vision, it has long been deemed to be a challenging task due to huge intra-class variations and inter-class ambiguities of scene images. Scene categorization can be applied to various tasks, such as image retrieval, human-computer interaction and intelligent robotics.

Appropriate representations are thought to be the most essential problem in scene categorization. To obtain suitable scene representations, researchers have made various attempts [1]–[4], e.g. directly utilizing global features, aggregating local features, and exploring mid-level visual representations. Although these methods have improved the performances of scene categorization so well, they rely on human to design features whose applicability is restrictive as image representations. Therefore, when the contents of the scene images are complicated, the categorization performances are degraded.

Recently, remarkable successes have been made by convolution networks (CNNs) in various visual tasks [5]–[7]. To obtain richer higher-level semantic features from images, deep convolution networks process raw data via a sequence of computational units, so that inputs can be transformed into some intrinsic representations. Many methods based on CNNs have been proposed for scene categorization [8]–[10], which significantly improve the

The associate editor coordinating the review of this manuscript and approving it for publication was Haluk Eren.

effect of scene categorization. After being trained on large-scale image datasets, such as ImageNet [11] and Places [7], CNNs can generate deep features which represent semantic information of images.

Although CNNs are powerful in capturing high-level features of objects, the representations are sometimes unclear when used to classify scene images directly. Since most of the current methods regard the scene as a set of multiple objects [12], and the description of contextual semantic relationship between the objects and the surroundings is insufficient, which affects the performance of scene categorization.

To this end, this paper uses the context-based saliency detection algorithm [13] to label the visually sensitive regions which contain the dominant objects and the partial background regions to express context information of scene images. Furthermore, the CNNs which are pre-trained on large-scale scene-centric dataset Places are combined to construct a scene categorization model based on deep visually sensitive features. By this model, we can capture more spatial structures of scene images and effectively overcome the limitations of using objects to classify simply, thus obtaining good categorization performances.

The rest of this paper is organized as follows. Section 2 presents an overview of the proposed method. Section 3 addresses the detailed categorization model. Section 4 analyzes and evaluates the results. Finally, we conclude this paper in Section 5.

II. RELATED WORK

So far, in order to obtain discriminative representations of scene, a number of scene categorization methods have been proposed.

A. UTILIZING GLOBAL FEATURES

Originally, scene images are represented by directly extracting the global features of the images [14]. Two representative features are Gist [15] and CENTRIST [3]. Gist uses spectral and roughly local information to capture the dominant spatial structures of the scene images and create an image representation [15]. CENTRIST uses a local structure distribution to capture the general structural characteristics of the scene images [3].

Although it is usually efficient to calculate the global features of scene images, the global features can only capture the global low-level information in scene images, while many details in scenes are ignored. As a result the global features are limited to deal with large variations in scene images, leading to poor performance in the case of complex scene categories.

B. AGGREGATING LOCAL FEATURES

Compared with the direct extraction of global features, local features are more robust against image content and scale variations. The methods generally represent images by aggregating local features, such as SIFT [16] and HOG [17]. The methods commonly used local feature clustering are BoVW [2], Fisher vectors [18] and sparse coding [19].

Compared with global features, the statistical results of local features have certain semantic information. They have significantly improved the categorization performance [20], [21]. However, since the local features are created from local image patches, they lack correlation information of local image semantics. Furthermore, because the scene category is a high-level visual concept, there is a semantic gap between the high-level semantics and low-level features of the scene images, which lacks sufficient discrimination for the judgment of the scene category. Thus, the method using local features cannot fundamentally improve the performance of categorization.

C. MID-LEVEL VISUAL REPRESENTATIONS

To overcome the semantic gap between low-level features and high-level information, mid-level features that reflect objects and other parts of scenes are widely used to classify the scene images. The mid-level features carry more meaningful semantic information and can better reflect the effective information that determines the scene category in the scene images.

Doerschet et al. used the mean-shift algorithm to seek the discriminative modes in the distribution space of image blocks and create the mid-level scene image representations [20]. In [21], Izadinia et al. used a joint learning procedure to learn the appearances and layout of objects simultaneously from scene categories. Parizi et al. proposed a latent variable model to represent the scene as a set of reconfigurable region models [22]. Singh et al. used iterative and cross-validation methods to obtain discriminative image patches as the mid-level representations of images [23].

During the period, people also found that most of the scenes contain some representative objects or regions, and sometimes the objects and layouts could be judged by the extraction and combination of objects. In [24], Quattoni and Torralba used manual segmentation to obtain partial regions of scene images and utilize these regions as valid information of scene categorization, demonstrating the effectiveness of the method. Furthermore, many effective methods using region information of scene images are proposed [4], [24]–[26].

The mid-level visual expressions overcome the disadvantages of low-level features lacking semantic description of images and high-level features modeling difficultly [27]. However, middle-level features are difficult to construct and generalize. Therefore, its development is limited.

D. EXPRESSION OF CNNs FEATURES

At present, the scene categorization algorithm based on deep learning has made a significant breakthrough, and more and more researchers use convolution neural networks to solve the problem of scene categorization [5], [7], [8]–[10].

To obtain better categorization results, Donahue et al. used CNNs trained on the dataset ImageNet for scene

categorization [5]. Zhou et al. proposed a large-scale scene-centric dataset Places to train CNNs, which significantly improves the performance of scene categorization [7]. Bai et al. proposed the transfer knowledge learning by CNNs from object-centric dataset to construct scene-specific object models for scene categorization [10]. Herranz et al. used scale-specific CNNs and multi-scale architectures to learn the knowledge of objects and scenes for classifying scene images [8].

Although the above methods have achieved obvious categorization effects, most of them just use a single layer of CNNs [5], [7]–[9]. The different layers of CNNs can generate different abstract patterns, but the single layer features cannot effectively utilize the multiple layer abstraction and expression of scene images. Therefore, in this paper, we propose to obtain scene images representations by concatenate multiple layers deep visually sensitive features of CNNs for scene categorization.

III. FRAMEWORK OF THE PROPOSED METHOD

In order to adapt to the diversity of scene images, we utilize the context-based saliency detection algorithm and deep convolution networks to build a scene categorization model, so that it can represent the deep intrinsic characteristics of the scene images.

A. VISUALLY SENSITIVE REGION DETECTION

Regions that have a major impact on visual judgment are called visually sensitive regions, which can be extracted by the context-based saliency detection algorithm [13]. The visually sensitive regions extracted fully consider both locally and globally distinctiveness at multi-scales. The salient objects together with the parts of regions that surround them can throw light on the meaning of images. The salient regions perfectly reflect the context between the objects and the surrounding regions in the scene, and filter out some repeated texture information.

According to principles of human visual attention, a pixel is considered salient if the appearance of the patch centered the pixel is distinctive with respect to all other image patches. In addition, positional distance between patches is also an important factor.

A dissimilarity measure between a pair of patches is defined as:

$$d(p_i, p_j) = \frac{d_c(p_i, p_j)}{1 + c \cdot d_p(p_i, p_j)} \tag{1}$$

where pixel i is center pixel of patches p_i. $d_c(p_i, p_j)$ is the Euclidean distance between the vectorized patches p_i and p_j in CIEL^*a^*b color space. $d_p(p_i, p_j)$ is the Euclidean distance between the positions of patches p_i and p_j.

For every patch p_i at single-scale, we search for the M most similar patches F_{fc7} in the image, according to Equation (1). A pixel i is salient when $d(p_i, p_j)$ is high $\forall m \in [1, M]$. The single-scale saliency value of pixel i at scale r is defined as:

$$S_i^r = 1 - \exp\left\{-\frac{1}{M}\sum_{m=1}^{M} d(p_i^r, p_m^r)\right\} \tag{2}$$

In order to make the detected saliency at multiple scales, it is necessary to calculate the mean of saliency at different scales:

$$\overline{S}_i = \frac{1}{K}\sum_{r \in R} S_i^r \tag{3}$$

where K denotes the number of scales, R denotes the scale space.

In addition, it is necessary to modify the saliency according to the context of images, so that the regions with different distances from the significant objects have different saliency. The saliency of a pixel is redefined as:

$$\widehat{S}_i = \overline{S}_i(1 - d_f(i)) \tag{4}$$

where $d_f(i)$ denotes the Euclidean positional distance between pixel i and the closest focus of attention pixel, normalized to the range [0,1]. The saliency of interesting background in the neighborhood of the salient objects will be increased.

Figure 1 shows an example of visually sensitive region detection. The brightness value in Figure 1(b) is the visual sensitivity of the position. It can be seen that the sensitivity of the background area varies with the degree of closeness to the objects.

(a) (b)

FIGURE 1. Example of visually sensitive region detection. (a) Original images. (b) Visually sensitive region detection images.

B. CONSTRUCTING VISUALLY SENSITIVE REGION ENHANCED IMAGES

Although the original images contain comprehensive scene information, they cannot distinguish between valid information and invalid information. Although the visually sensitive region detection images can distinguish different significant regions, they may lose some detail information that can assist in expressing image contents. In order to solve this problem, we propose to superimpose the original images with the corresponding visually sensitive region detection images to obtain the visually sensitive region enhancement images.

The superimposed form is as follows:

$$f_e(i, j) = f_s(i, j). * f(i, j) \tag{5}$$

$f_e(i, j)$ denotes the visually sensitive region enhancement image, $\cdot^*$ denotes dot product operation, $f_s(i, j)$ denotes the visually sensitive region detection image, $f(i, j)$ denotes the

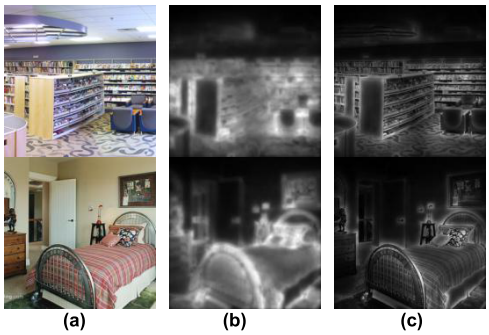


FIGURE 2. Visually sensitive region enhancement. (a) Original images. (b) Visually sensitive region. (c) The result of detection images product (a) and (b).

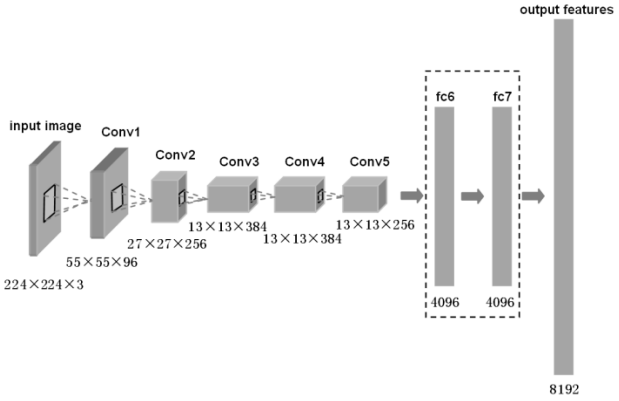


FIGURE 3. Schematic diagram of generating deep convolution features utilized the full-connection layers of Alexnet CNN model.

original image. They are normalized to the range [0, 1] before superimposing.

Figure 2 shows visually sensitive region enhancement. It can be seen from Figure 2(b) that the visual sensitivity of different regions in the scene is different, and some repeated textures and insignificant region information in the scene are effectively suppressed, such as lights of the library, doors of the bedroom. The superimposed image supplements some detail information (transition regions between sensitive regions and insensitive regions) on the basis of preserving visual sensitivity.

C. DEEP VISUALLY SENSITIVE FEATURES

The original image, the visually sensitive region detection image and the visually sensitive region enhancement image are respectively input into an existing AlexNet CNN model, which consists of convolution layers, pooling layers and full-connection layers to extract deep convolution features. The model has been pre-trained on the large-scale scene dataset Places. Because the deep features from different layer of CNNs correspond to different levels of abstraction of input images, we use representations from multiple layers of CNNs for classifying scene images.

Figure 3 shows the schematic diagram of generating deep convolution features utilized the full-connection layers of Alexnet CNN model. The 4096-dimensional output features of the $FC7$ and $FC6$ are concatenated to generate deep convolution features of input image. The calculation formula is as follows:

$$F_c = [F_{fc7}, F_{fc6}] \tag{6}$$

where F_c denotes the deep convolution features of input image, F_{fc7} is the output features of $FC7$, F_{fc6} is the output features of $FC6$.

After that, to mark different visually sensitive regions and preserve some details information, the deep convolution features of the original image, the visually sensitive region detection image and the visually sensitive region enhancement image of the same image are concatenated, generating deep visually sensitive features. The form is as follows:

$$F_{VS-CNN} = [F_c, F_{c-s}, F_{c-e}] \tag{7}$$

where F_{VS-CNN} denotes deep visually sensitive features of the image, F_c, F_{c-s} and F_{c-e} are the deep convolution features of the original image, the deep convolution features of visually sensitive region detection image, and the deep convolution features of visually sensitive region enhancement image, respectively.

The deep visually sensitive features extracted from training images of Benchmark datasets are input into the SVM to train a linear SVM classifier for each scene category. Since the extracted features contain not only the context semantic relationship between objects and surroundings in images, but also the deep intrinsic characteristics of the scene images, the obtained model is called Visually Sensitive CNN model (VS-CNN).

IV. EXPERIMENTS

A. DATASETS AND SETTINGS

For assessing the proposed method, we conduct extensive experiments on five challenging benchmark scene datasets that are public available online, i.e. LabelMe(OT) [15], UIUC-Sports(SE) [12], Scene-15(LS) [2], [15], [27], MIT67(IS) [23] and SUN [28]. Figure 4 shows the partial image of each scene dataset. For comparing with the similar algorithms, the standard train/test split for the different dataset is adopted in experiments. In addition, since the number of images in different categories of each dataset is various greatly, in order to maintain the balance of training data, the same number of samples is randomly selected for each categories of the same dataset. The experiments are performed for 10 times with the average accuracy reported.

. LabelMe(OT): This dataset contains 2688 color outdoor images in eight different categories: MITcoast (360 images), MITforest (328 images), MIThighway (260 images), MITinsidecity (308 images), MITmountain (374 images), MITopencountry (410 images), MITstreet (292 images), and MITtallbuilding (356 images). The size of images is 256 × 256 pixels. We randomly select 200 images from each category for training and the rest for testing each time, respectively.

FIGURE 4. The partial images of each dataset. (a) The partial images of dataset OT. (b) The partial images of dataset SE. (c) The partial images of dataset LS. (d) The partial images of dataset IS. (e) The partial images of dataset SUN.

. **UIUC-Sports(SE):**This dataset contains 1579 color images in eight sports event categories: badminton (200 images), bocce (137 images), croquet (236 images), polo (182 images), rock climbing (194 images), rowing (250 images), sailing (190 images), and snowboarding (190 images). The size of images is different. We randomly select 70 images from each category for training and another 60 images for testing each time, respectively.

. **Scene-15(LS):**This dataset contains 4485 indoor and outdoor scene images of 15 different categories, eight of which are the same as the LabelMe dataset. The rest categories are bedroom (216 images), CALsuburd (241 images), industrial (311 images), kitchen (210 images), living room (216 images), PARoffice (215 images), and store (315 images). We randomly select 100 images from each category for training and the rest for testing each time, respectively.

. **MIT67(IS):** This is a challenging dataset of 15620 indoor scene images which contains 67 different categories, such as classroom, library, children-room and so on. We randomly select 80 images from each category for training and another 20 images for testing each time, respectively.

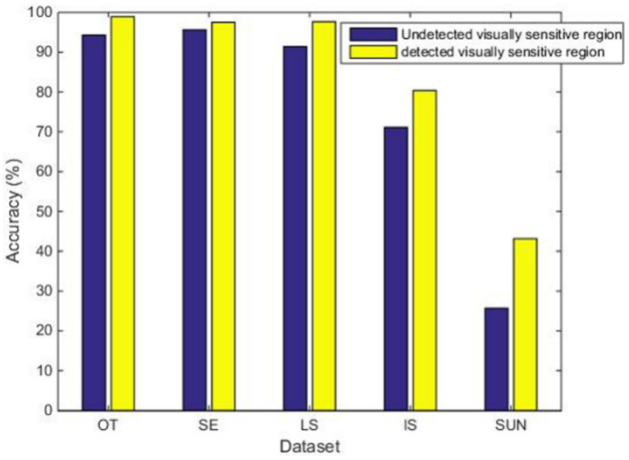


FIGURE 5. Comparison of utilizing visually sensitive region detection.

. **SUN:** The dataset is the most challenging dataset for scene categorization. It contains 397 scene categories and each has at least 100 images. We randomly select 50 images from each category for training and another 50 images for testing each time, respectively.

B. ABLATION STUDIES

In this paper we extract the deep visually sensitive features from full-connection layers of Alexnet for representing images, and train a SVM classifier for scene categorization. Here, we first investigate the performance of utilizing the visually sensitive region detection on five scene datasets. Obtained results are shown in Figure 5.

From Figure 5, it can be observed that the categorization performance of dataset SUN is improved obviously. The main reason is that the contents of scene images in this dataset have rich context relationship. It also shows that the proposed model has better performance. Furthermore, the categorization accuracy for the datasets which contain indoor scenes has obvious improvement as well, such as datasets LS and IS. The main reason is that the number of objects is large and the interrelationship is complex. Therefore, the description of the context relationship of salient objects in scene images can fully represent scene properties.

However, the improvement of categorization performance on dataset SE is not obvious. Because the sports event scene is composed of persons and surroundings together, and persons who are not the principal objects for distinguishing scenes can appear in different scenes. The discrimination of the scene category is mainly determined by the relationship between the character action and the surrounding environment.

To evaluate the performance of the proposed categorization model which utilizes concatenate features, the VS-CNN model with output features from the $FC6$, $FC7$ and $FC6 + FC7$ are compared in terms of accuracy (%) on five datasets, respectively. The comparison results are shown in Table 1.

From Table 1, we can see that when adopting concatenate features from the $FC6$ and $FC7$, the best categorization result

TABLE 1. Classification results achieved in terms of accuracy (%) by utilizing different full-connection layers at the output.

	*FC*6	*FC*7	*FC*6+*FC*7
OT	98.81	98.62	**98.90**
SE	96.67	96.46	**97.50**
LS	97.52	97.18	**97.65**
IS	79.61	78.36	**80.37**
SUN	42.04	41.22	**43.14**

TABLE 2. Classification results achieved in terms of accuracy (%) by utilizing different input images at the output.

	VSR	VSE	VSR+ORI	VSE+ORI	VSR+VSE+ORI
OT	80.69	89.15	95.59	98.71	**98.90**
SE	66.88	79.79	94.58	96.46	**97.50**
LS	70.12	84.65	92.56	97.52	**97.65**
IS	22.61	45.37	69.93	79.63	**80.37**
SUN	12.34	19.32	30.86	41.55	**43.14**

TABLE 3. Confusion matrix of dataset OT.

Categorization	1	2	3	4	5	6	7	8	Recall(%)
MITcoast（1）	159	1	-	-	-	-	-	-	99.4
MITforest（2）	-	128	-	-	-	-	-	-	**100**
MIThighway（3）	-	-	60	-	-	-	-	-	**100**
MITinsidecity（4）	-	-	-	108	-	-	-	-	**100**
MITmountain（5）	1	-	-	-	173	-	-	-	99.4
MITopencountry（6）	4	-	-	-	1	205	-	-	97.6
MITstreet（7）	-	-	-	-	-	-	92	-	**100**
MITtallbuilding（8）	-	-	-	-	-	-	-	156	**100**
Precision（%）	96.9	99.2	**100**	**100**	99.4	**100**	**100**	**100**	

FIGURE 6. Misclassification examples in the dataset OT.

is obtained, especially on dataset SUN. Analyzing performance layer-by-layer shows that features from *FC*7 generalize worse than features from *FC*6. It also reveals that much of the CNN's representational power comes from *FC*6, rather than from *FC*7. The experimental results demonstrate that the accuracy of scene categorization will be limited when only the single layer features are used and the proposed VS-CNN model can greatly improve the categorization performance by effectively fusing the multiple layer features.

Next, to evaluate the impact of different input images of CNNs on categorization performance, we compare and analyze experimental results of only utilizing visually sensitive region detection image (VSR), visually sensitive region enhancement image (VSE), visually sensitive region detection image together with original image (VSR+ORI) and visually sensitive region enhancement image together with original image (VSE+ORI), respectively. Obtained results are shown in Table 2.

By comparing the results in Table 2, we have the following points. First, the performance of VSR+VSE+ORI is generally better than the other four types input images and the performance of VSR is the worst. The reason is that VSR removed some visually insensitive information of scene images, which demonstrates that it is beneficial to use ORI data for scene categorization tasks, especially when the content of scene images is quite complicated. Second, the proposed model gained the greatest improvement for dataset IS than the other four datasets. More specifically, the categorization precision of VSR+VSE+ORI than that of VSR is significantly increased by nearly 60% on dataset IS. The proposed VS-CNN model utilizing superimposing of the different input images significantly outperformed the other methods mainly because it takes advantage of the different details of images, which can properly maintain the structure properties of each feature and adequately exploit the complementary information between the different features.

C. PERFORMANCE EVALUATION

To demonstrate the effectiveness of the proposed model for scene categorization, we compare and analyze experimental results on three datasets using precision and recall criterion.

The confusion matrix of dataset OT is shown in Table 3. It can be observed that the precision and recall can achieve 100% on 'MIThighway', 'MITinsidecity', 'MITstreet' and 'MITtallbuilding' utilizing proposed model. The rest of the categories can also achieve good categorization results. The most mistakes are that the 'MITopencountry' is misclassified into the 'MITcoast'.

The misclassification examples of Table 3 are partially shown in Figure 6. The first image is to misclassify 'MITopencountry' into 'MITmountain', the second image is to misclassify 'MITcoast' into 'MITforest', and the last two images are to misclassify 'MITopencountry' into 'MITcoast'. From Figure 6, it can be seen that the reason for misclassification is mainly that the images contain features of misclassified scene category.

The confusion matrix of dataset SE is shown in Table 4. It can be observed that the precision and recall can achieve 100% on 'Rock Climbing', 'rowing', 'sailing' and 'snowboarding'. 'Bocce' and 'croquet' are the two most confusing categories, in which bocce has the lowest precision. The main reason is that the surroundings of the two scene categories are very similar and the discrimination of the two scenes mainly depends on the characters behavior. The character actions of the two scenes are very close, so it is easy to misclassify.

The misclassification examples of Table 4 are partially shown in Figure 7. The first image is to misclassify 'bocce'

TABLE 4. Confusion matrix of dataset SE.

Categorization	1	2	3	4	5	6	7	8	Recall(%)
badminton (1)	59	1	-	-	-	-	-	-	98.3
bocce (2)	1	55	3	1	-	-	-	-	91.7
croquet (3)	-	5	55	-	-	-	-	-	91.7
polo (4)	-	-	-	60	-	-	-	-	**100**
RockClimbing (5)	-	-	-	-	60	-	-	-	**100**
rowing (6)	-	-	-	-	-	60	-	-	**100**
sailing (7)	-	-	-	-	-	-	60	-	**100**
snowboarding (8)	-	-	-	-	-	-	-	60	**100**
Precision (%)	98.3	90.2	94.8	98.4	**100**	**100**	**100**	**100**	

FIGURE 7. Misclassification examples in the dataset SE.

into 'badminton'. The reason is that the throwing action in 'bocce' is similar to the catching action in 'badminton'. The second image is to misclassify 'bocce' into 'polo', that's due to the green plants in 'bocce' are similar to that in 'polo'.

The partial categorization results of dataset IS are shown in Figure 8. It can be seen that the categorization results of 'bowling', 'cloister', 'greenhouse' and 'restaurant' reach 100%, while the categorization result of the 'deli' category is only about 30% at the lowest. It shows that the proposed model has limited ability to represent 'deli' category.

The confusion matrix of dataset LS is shown in Table 5. It can be observed that the precision and recall can achieve 100% on 'CALsuburb' and 'MIThighway'. 'MITopencountry' is most easily misclassified to 'MITcoast'. As shown in the first two images of Figure 9, the texture information of 'MITopencountry' is similar to that of 'coast'. The reason for the misclassification of the last two images is that the tall buildings, clouds, and streetlights are similar to the buildings of 'industrial' scene.

D. COMPARISON TO STATE-OF-THE-ART METHODS

To demonstrate the effectiveness of the proposed method, we compare it to state-of-the-art methods on five challenging benchmark datasets for scene categorization. All methods used for comparison follow the same protocols described in the experiment setting subsection. The evaluation criterion for these datasets is the average categorization accuracy over the 10 splits. Comparison results of the proposed method to state-of-the-art approaches on five datasets are shown in Table 6.

From the result, we can see that approaches based on deep features give superior performances to traditional ones. For the dataset OT, deep networks trained by using the scene-centric database Places show better performances than the one trained on the object-centric database ImageNet. The deep features of the proposed method can extract richer semantic information and represent the association of the semantic information.

For the dataset SE, what is different is that the proposed method only slightly outperforms ImageNet-CNN [11] trained on the dataset ImageNet and Places-CNN [7] trained on the dataset Places. It is mainly due to the strong interference of the actions of the characters in the sports scenes, and the sensitive regions often include the people in scenes. The richness of the sports movement has a certain influence on the category determination. Furthermore, although Scene-specific objects+Structure features [10] utilize the structure

TABLE 5. Confusion matrix of dataset LS.

Categorization	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Recall (%)
Bedroom (1)	111	-	-	-	5	-	-	-	-	-	-	-	-	-	-	95.7
CALsuburb (2)	-	141	-	-	-	-	-	-	-	-	-	-	-	-	-	**100**
Industrial (3)	1	-	205	-	1	-	-	-	3	-	-	-	-	-	1	97.2
Kitchen (4)	-	-	-	107	1	-	-	-	-	-	-	-	-	1	1	97.3
Livingroom (5)	6	-	-	3	177	-	-	-	-	-	-	-	-	3	-	93.7
MITcoast (6)	-	-	-	-	-	254	1	-	-	-	5	-	-	-	-	97.7
MITforest (7)	-	-	-	-	-	-	225	-	-	1	2	-	-	-	-	98.7
MIThighway (8)	-	-	-	-	-	-	-	160	-	-	-	-	-	-	-	**100**
MITinsidecity (9)	-	-	3	-	1	-	-	-	204	-	-	-	-	-	-	98.1
MITmountain (10)	-	-	-	-	-	1	-	-	-	273	-	-	-	-	-	99.6
MITopencountry (11)	-	-	-	-	-	10	1	-	-	1	298	-	-	-	-	96.1
MITstreet (12)	-	-	2	-	1	-	-	-	1	-	-	188	-	-	-	97.9
MITtallbuilding (13)	-	-	7	-	-	-	-	-	4	-	-	-	245	-	-	95.7
PARoffice (14)	-	-	-	-	-	-	-	-	-	-	-	-	-	115	-	**100**
Store (15)	-	-	-	1	1	-	-	-	-	-	-	-	1	-	212	98.6
Precision(%)	94.1	**100**	94.5	96.4	94.7	95.9	99.1	**100**	96.2	99.4	97.7	**100**	99.6	96.6	99.1	

TABLE 6. Comparison to state-of-the-art approaches on five datasets (%).

	OT	SE	LS	IS	SUN
LGF[1]	-	88.52	85.80	-	-
O2C kernels[31]	-	86.02	88.81	39.85	-
ISPR+IFV[35]	-	92.08	91.06	68.50	-
ImageNet-CNN[11]	92.83	94.42	84.23	56.79	42.61
Places-CNN[7]	94.30	94.12	90.19	68.24	**54.32**
Hybrid-CNN[7]	-	94.22	91.59	70.80	53.86
RF-CNNs[32]	-	94.86	-	72.35	-
S^2ICA[33]	-	95.80	93.1	71.20	-
DDSFL+Caffe[9]	-	96.78	92.81	76.23	-
DAG-VggNet19[38]	-	-	-	77.50	-
Scene-specific objects+Structure features[10]	-	96.93	-	74.35	-
SDO+ fc features[36]	-	-	95.88	-	-
OTC[39]	-	-	84.40	47.33	34.56
M-CNN[34]	-	-	87.50	78.90	42.40
LRML-PCDM[29]	-	76.97	65.82	-	-
DGSK[30]	-	-	92.30	75.10	-
DFDF[37]	87.1	76.3	-	-	-
VS-CNN (Ours)	**98.90**	**97.50**	**97.65**	**80.37**	43.14

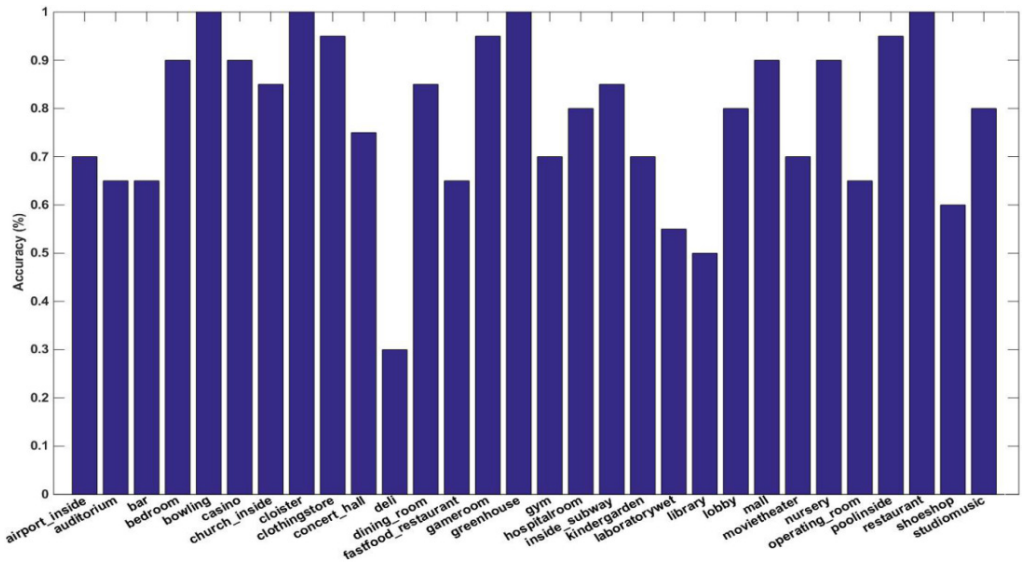


FIGURE 8. The partial categorization results of dataset IS.

and objects information of the scene, the deep visually sensitive features obtained by using the context information of the visually sensitive regions are obviously more effective.

For the dataset LS, SDO + fc features [36] exploit the correlations of object configurations among different scenes by the co-occurrence pattern of all objects across scenes to choose representative and discriminative objects which enhance the discriminating ability of inter-class. Moreover, the method represents the image descriptors with the occurrence probabilities of discriminative objects in image patches to eliminate the negative effects of common objects. Although the method considers the correlation between the objects in the scene, it is still limited to the simple objects, and the surrounding background area adjacent to the object is not considered, so the effectiveness is limited.

For the dataset IS, the accuracy of Places-CNN pre-trained on the dataset Places exceeds that of ImageNet-CNN pre-trained on the dataset ImageNet by nearly 12%. The reason is that the spatial structure features of networks trained by the scenes are more effective for scene categorization. Furthermore, because the variances of indoor scenes are greater, the randomness is also greater. Therefore, compared to outdoor scenes, the advantage of the proposed method is obvious. The proposed method obtained better results than the Hybrid-CNN method, in which image features are extracted from CNNs pre-trained on a dataset obtained by combining datasets ImageNet and Places. Our method only uses the convolution neural network pre-trained by the Places dataset, and combines the context information of the visually sensitive region of images to obtain higher accuracy than the Hybrid-CNN method by nearly 10%.

However, for the dataset SUN, which is the largest dataset for scene categorization, the performance of the proposed method is lower than Places-CNN and Hybrid-CNN.

TABLE 7. The parameter training times(s) of the SVM classifier.

Dataset	OT	SE	LS	IS	SUN
Parameter training times	49.01	14.68	79.24	1432.62	9348.71

FIGURE 9. Misclassification examples in the dataset LS.

The reason is that some regions that cannot represent the content properties of scenes are extracted as visually sensitive regions in scene images, resulting in the impact on classification accuracy. In future work, we will do more in-depth research on the utilizing of context information. But it can be seen that the proposed model still has obvious advantages compared that of the two methods on the other datasets. It also shows that the proposed model has good adaptability under multiple datasets.

In addition, Table 7 shows the parameter training times of the SVM classifier in the proposed VS-CNN model on five datasets. The processor is Intel(R) Core(TM) i7-4790, CPU @ 3.60GHz. The experiments are performed for 10 times for reporting their average time.

V. CONCLUSION

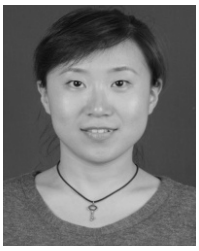
In this paper, a scene categorization model based on deep visually sensitive features is proposed for scene categorization. To this end, we utilize the context-based saliency

detection to obtain the visually sensitive regions of scene images, superimposed with the corresponding original image to obtain the visually sensitive region enhancement image. Furthermore, to utilize the complementary information of different layers of deep convolution neural networks, we extract multiple types of features from images based on its hierarchical structure and concatenate them as the representations of scene images.

Because the extracted features by the proposed model simultaneously describes the object information and the context semantic information between the object and its surrounding scenes in images, the proposed model can obtain the visual expression of the scene images reasonably by combining different visual sensitivities with the multi-layer deep convolution features. Extensive experiments are conducted on the five benchmark scene datasets i.e. LabelMe, UIUC-Sports, Scene-15, MIT67 and SUN. Experiment results demonstrate that the proposed method is effective for scene categorization and can achieve superior results to most of state-of-the-art approaches.

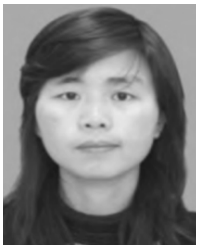
REFERENCES

[1] J. Zou, W. Li, C. Chen, and Q. Du, "Scene classification using local and global features with collaborative representation fusion," *Inf. Sci.*, vol. 348, pp. 209–226, Jun. 2016.
[2] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, Jun. 2006, pp. 2169–2178.
[3] J. Wu and J. M. Rehg, "Centrist: A visual descriptor for scene categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1489–1501, Aug. 2011.
[4] K. A. de Souza Gazolli and E. O. T. Salles, "Exploring neighborhood and spatial information for improving scene classification," *Pattern Recognit. Lett.*, vol. 46, pp. 83–88, Sep. 2014.
[5] J. Donahue *et al.*, "Decaf: A deep convolutional activation feature for generic visual recognition," in *Proc. Int. Conf. Mach. Learn.*, Beijing, China, Jan. 2014, pp. 647–655.
[6] D. Yoo, S. Park, J.-Y. Lee, and I. S. Kweon, "Multi-scale pyramid pooling for deep convolutional representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Boston, MA, USA, Jun. 2015, pp. 71–80.
[7] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 487–495.
[8] L. Herranz, S. Jiang, and X. Li, "Scene recognition with CNNS: Objects, scales and DATASET bias," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 571–579.
[9] Z. Zuo, G. Wang, B. Shuai, L. Zhao, and Q. Yang, "Exemplar based deep discriminative and shareable feature learning for scene image classification," *Pattern Recognit.*, vol. 48, no. 10, pp. 3004–3015, 2015.
[10] S. Bai, "Scene Categorization Through Using Objects Represented by Deep Features," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 31, no. 9, Sep. 2017, Art. no. 1755013.
[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet lassification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1106–1114.
[12] L.-J. Li and L. Fei-Fei, "What, where and who? Classifying events by scene and object recognition," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Rio de Janeiro, Brazil, Oct. 2007, pp. 1–8.
[13] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1915–1926, Oct. 2012.
[14] P. Wang, J. Wang, G. Zeng, W. Xu, H. Zha, and S. Li, "Supervised kernel descriptors for visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 2858–2865.
[15] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.
[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Diego, USA, Jun. 2005, pp. 886–893.
[18] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *Int. J. Comput. Vis.*, vol. 105, no. 3, pp. 222–245, 2013.
[19] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 801–808.
[20] C. Doersch, A. Gupta, and A. A. Efros, "Mid-level visual element discovery as discriminative mode seeking," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 494–502.
[21] H. Izadinia, F. Sadeghi, and A. Farhadi, "Incorporating scene context and object layout into appearance modeling," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 232–239.
[22] S. N. Parizi, J. G. Oberlin, and P. F. Felzenszwalb, "Reconfigurable models for scene recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 2775–2782.
[23] S. Singh, A. Gupta, and A. A. Efros, "Unsupervised discovery of mid-level discriminative patches," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, Oct. 2012, pp. 73–86.
[24] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 413–420.
[25] N. Morioka and S. Satoh, "Building compact local pairwise codebook with joint feature space clustering," in *Proc. Eur. Conf. Comput. Vis.*, Heraklion, Greece, Sep. 2010, pp. 692–705.
[26] K. Gazolli and E. Salles, "A contextual image descriptor for scene classification," in *Proc. Online Trends Innov. Comput.*, 2012, pp. 66–71.
[27] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Diego, CA, USA, Jun. 2005, pp. 524–531.
[28] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 3485–3492.
[29] G. Sun, Y. Cong, Q. Wang, and X. Xu, "Online low-rank metric learning via parallel coordinate descent method," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Beijing, China, Aug. 2018, pp. 207–212.
[30] X. Sun *et al.*, "Scene categorization using deeply learned gaze shifting Kernel," *IEEE Trans. Cybern.*, vol. 49, no. 6, pp. 2156–2167, May 2018.
[31] L. Zhang, X. Zhen, and L. Shao, "Learning object-to-class kernels for scene classification," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3241–3253, Aug. 2014.
[32] S. Bai, "Growing random forest on deep convolutional neural networks for scene categorization," *Expert Syst. Appl.*, vol. 71, pp. 279–287, Apr. 2017.
[33] M. Hayat, S. H. Khan, M. Bennamoun, and S. An, "A spatial layout and scale invariant feature representation for indoor scene classification," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4829–4841, Oct. 2016.
[34] W. Yin, D. Xu, Z. Wang, Z. Zhao, C. Chen, and Y. Yao, "Perceptually learning multi-view sparse representation for scene categorization," *J. Vis. Commun. Image Represent.*, vol. 60, pp. 59–63, Apr. 2019.
[35] D. Lin, C. Lu, R. Liao, and J. Jia, "Learning important spatial pooling regions for scene classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 3726–3733.
[36] X. Cheng, J. Lu, J. Feng, B. Yuan, and J. Zhou, "Scene recognition with objectness," *Pattern Recognit.*, vol. 74, pp. 474–487, Feb. 2018.
[37] H. Shan, J. Zhang, and U. Kruger, "Framework of randomized distribution features for visual representation and categorization," *IEEE Trans. Cybern.*, to be published.
[38] S. Yang and D. Ramanan, "Multi-scale recognition with DAG-CNNs," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1215–1223.
[39] R. Margolin, L. Zelnik-Manor, and A. Tal, "OTC: A novel local descriptor for scene classification," in *Proc. Eur. Conf. Comput. Vis.*, Zurich, Switzerland: Springer, Sep. 2014, pp. 377–391.



JING SHI received the B.S. and M.S. degrees from the Institute of Automation and Information Engineering, Xi'an University of Technology, Xi'an, China, in 2006 and 2009, respectively. She is currently pursuing the Ph.D. degree with the Institute of Automation and Information Engineering, Xi'an University of Technology. Her research interests include pattern recognition, scene classification, and digital images processing.

HONG ZHU received the Ph.D. degree from Fukui University, Fukui, Japan, in 1999. She is currently a Professor with the Institute of Automation and Information Engineering, Xi'an University of Technology, Xi'an, China. Her research interests include image analysis, intelligent video surveillance, and pattern recognition.



SHUNYUAN YU received the Ph.D. degree from the Institute of Automation and Information Engineering, Xi'an University of Technology, Xi'an, China, in 2017. She is currently with the Institute of Electronic and Information Engineering, Ankang University. Her research interests include pattern recognition and digital images processing.

WENHUAN WU received the M.S. degree in computer application technology from Nanchang Hangkong University, Nanchang, China, in 2009. He is currently pursuing the Ph.D. degree with the School of Automation and Information Engineering, Xi'an University of Technology, Xi'an, China. He is also a Lecturer with the Hubei University of Automotive Technology, Shiyan, China. His research interests include computer vision, pattern recognition, and image processing.

HUA SHI received the Ph.D. degree from the Institute of Automation and Information Engineering, Xi'an University of Technology, Xi'an, China, in 2017. She is currently with the Institute of Sciences, Xi'an Technological University. Her research interests include stereo vision, pattern recognition, and digital images processing.