Design and Development of Scene Recognition and Classification Model Based on Human Pre-attentive Visual Attention

A R Kudus and C S Teh
Faculty of Cognitive Sciences and Human Development, University Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia

E-mail: 16020136@siswa.unimas.my

Abstract. Recent works on scene classification still utilize the advantages of generic feature of Convolutional Neural Network while applying object-ontology technique that generates limited amount of object regions. Human can successfully recognize and classify scene effortlessly within short period of time. By utilizing this idea, we present a novel approach of scene classification model that built based on human pre-attentive visual attention. We firstly utilize saliency model to generate a set of high-quality regions that potentially contain salient objects. Then we apply a pre-trained Convolutional Neural Network model on these regions to extract deep features. Extracted features of every region are then concatenated to a final features vector and feed into one-vs-all linear Support Vector Machines. We evaluate our model on MIT Indoor 67 dataset. The result proved that saliency model used in this work is capable to generate high-quality informative salient regions that lead to good classification output. Our model achieves a better average accuracy rate than a standard approach that classifies as one whole image.

1. Introduction
Scene classification problem has become more popular among researchers in the field of image recognition and computer vision as general [1]. Recognizing, classifying and understanding scene is a basic task in computer vision [2][3]. Scene recognition or classification is a process of organizing images and predicting the class category of the image environment [1][2]. This can be done through the process of supplying attributes information of the scene.

As for human, this is such an easy task. Human can successfully recognize and classify scene effortlessly within short period of time. Human recognizes scene’s category by observing objects present in the scene environment [3]. The existent objects would be the differentiation factor between two different scenes. Unfortunately, for a complex natural indoor scene which contains massive amount of different objects [3], humans are not able to observe and pay their attention on every object within short period of time [4], yet still able to recognize its category correctly. Indoor scene is consider to be hard to recognize compare to outdoor scene due to its different layouts, decorations and objects [3][5]. Observation process made by human visual attention and their eyes movement are very selective [6],[7]. In a glance, human only fixate and focus on some important objects/areas instead of fixating on the whole visual input entirely [6]. These objects are usually called salient proto-objects.
Recently, some scene recognition models have improved by applying object-ontology technique as their approach \([8][9][10][11][12]\). This approach is using objects appearance as discrimination characteristics of different scene images. For example, MetaObject-CNN by \([8]\) is using their region proposal technique to generate a set of discriminative patches of objects. These patches are used for the scene recognition process. Reference \([9]\) and \([10]\) have tried to reduce the problem of scale bias between scene dataset and object dataset by introducing multi-scale CNN architectures. Reference \([11]\) has used the posterior probability for the selection of discriminative objects which frequently presence in scene images. Scene recognition model by \([12]\) is adopting YOLOv2 \([13]\) to extract objects from scene images.

We believe that these methods do not specifically apply and focus on the concept of human visual attention when attending to objects and classifying scene images. Being able to mimic and apply the ability of human visual attention when recognizing scenes will greatly improve computer vision automation process. In order to achieve this assumption; firstly we have designed a scene recognition and classification model that apply object-ontology technique by utilizing the power of one of the earliest and successful saliency model to generate possible salient object region of a scene image. This approach is believe to simulate the selective process of human visual attention.

The goal of this study is to prove whether or not saliency model can help to provide useful/meaningful salient objects/regions and thus improve the scene classification accuracy.

In this paper, a novel scene recognition and classification model pipeline is presented. Every component and its process is explained in details. Experiment and benchmarking testing are conducted to validate and evaluate the model’s performance on classification task will also be presented.

2. Related Works

2.1. Scene Classification

The problem of recognizing and classifying scene is very significant and important in computer vision \([14]\). Scene classification, which can also be referred as scene recognition or scene analysis, is a long-standing research problem in computer vision. It is a process of assigning a label either as ‘beach’, ‘bedroom’, or simply ‘indoor’ or ‘outdoor’ to an input image, based on the image’s contents \([14][2]\).

For past two decades, some scene classification models have been proposed in many studies. Basically, they are based on local or global approach. For example, GIST \([15]\) uses the global scene descriptor and spatial pyramids representation of SIFT \([16]\) is detecting feature and identifying the position of the main points before using neighborhood gradient. Reference \([17]\) and \([18]\) proposed a model that try to capture mid-level information from deformable parts. Unfortunately, all these methods are based on low-level feature, which focus only on the statistical values of the image without the high-level semantic information resulting fail to provide strong discriminative characteristics \([2]\). These methods faced huge semantic gap between image representations and image recognition goals. Low-level image representation makes model to do more work in order to achieve higher-level recognition goals since low-level image representation has a weak connection between features and image description \([2]\).

The recent scene recognition models have tried to reduce the problem of ‘semantic gap’ between low-level and high-level image representation by applying object ontology technique, where objects as discriminative features representing scene image \([1][2][3][20]\). Some of the models implemented paradigms that able to generate unlimited amount of object regions. This approach is important for scene classification task \([20]\) and it is able to avoid unnecessary information in scene images \([3]\). Appearance and association of objects in an image will increase the chance of image to be classified to a certain category \([1]\).

A complex indoor scene image contains several different objects and concepts \([20]\), not all objects are essential for the process of classification \([3]\). Processing all available objects in indoor scene images will increase feature complexity and increase computational time. Therefore, processing only essential objects is important to reduce computational time and still perform well on the classification.
2.2. Convolutional Neural Network (CNN)

CNN is one of Deep Learning methods that frequently used to solve object recognition and classification problem [19][21]. CNNs are able to directly learn high-level features from raw data with multi-layer hierarchical transformation for image processing. This makes CNN able to learn discriminative deep features and it has replaced the low-level hand-crafted features in image processing and computer vision [19]. Pre-trained CNN models are the existence CNN architecture which has been trained on large-scale ImageNet database and can be reused to train on different dataset for different problem. Pre-trained CNN model is very flexible and stable to be applied as feature extractor [14] that extracts and generates feature vectors. The extracted features can be used to train and test Support Vector Machine Classifier.

2.3. Human Visual Attention

Visual attention has been usually compared to a virtual spotlight [22] and normally known as the process of selection what to observe by human eyes [23]. In the process of selective visual attention, human rapidly move their gaze towards objects of interest in their visual environment [6][7]. Selective visual attention has limited capability to process all visual information [4]. There is not enough attention for human eyes to focus on all objects/areas in their visual field at a time and collect all the information, therefore, it is important for them to choose and process only useful objects/areas [24]. Reference [14] explained that there is a lot of objects in human visual field especially in the environment of indoor scene and human eyes cannot simultaneously process all of it [4][6][7].

Viewed scenes are usually categorized by human according to different objects appear in the scene [3]. Unfortunately, since visual capacity is very limited to process all objects at once [4] in scene recognition, hence human perception tends to firstly attend to the most salient objects in the scene which guided by bottom-up influences. There are usually some parts of an image that particularly attract human’s attention. Human attention chooses a small portion of the visual inputs [4]. Reference [25] decade ago explained that, human attention temporarily gathers the volatile proto-objects coherent representations of attended objects before yield to scene recognition.

2.4. Saliency Model Conception

One of the most early representation of human visual attention, Saliency Toolbox [26], are biologically inspired and based on bottom-up computational framework [24], where the bottom-up influences are referring to multiple low-level visual such as color, intensity, orientation, texture and motion extracted from the input image at multiple scales [6]. Saliency Toolbox is actually built as one of the computational saliency models that used to predict the movements of human eyes on image. It works by attending to salient proto-objects within natural scene image. This saliency model will be implemented to identify meaningful objects in a scene because human visual attention system is generally unconsciously guided by low-level features of visual scene [6][24].

Saliency Toolbox [26] is operating based on three important operations: (i) winner-take-all, to identify the most salient image location, (ii) the argmax function, to determine the feature map which contain highest score of saliency contribution, and (iii) spreading process of activation, spreading the activation over the attended proto-object at the attended location.

3. Model Architecture

In this section, we present three main operation components of our approach for scene classification model. As illustrated in figure 1 step (a), our model implements saliency model to generate salient region proposals. The model is built on top of a pre-trained deep Convolutional Neural Network model that performs feature extraction process in step (c) and (d). Classification process is carried out by Multi-class Linear SVMs classifiers in step (e).
3.1. Generating Salient Region Proposals
We agreed that natural scene can be well classified using objects, but in a glance human visual attention are not able to attend to all objects and process every information. Therefore, we proposed to apply Saliency Toolbox Model in [26] to our model as the first step to generate salient region proposals. The application of Saliency Toolbox simulates the selective process of human visual attention when recognizing and classifying scene images.

Figure 1 illustrates Saliency Toolbox is applied first in step (a) and (b). It attends and generates $n$ amount of most salient region proposals by cropping the regions from the original image. These salient region proposals are resized to 224-by-224 image size. These regions are then feed into pre-trained CNN model. This is because pre-trained CNN model used in our model only receive input image with size 224-by-224 dimension [27].

3.2. Extracting and Concatenating Deep Features
GoogLeNet CNN pre-trained model [27] is selected as a generic feature extractor for all salient region proposals generated by Saliency Toolbox. This process is performed in step (c) and (d) as shown in figure 1. All salient region proposals are feed into GoogLeNet one by one. The salient regions are propagated through all the layers in the pre-trained CNN model. GoogLeNet will perform the learning process and produce deep features of a region proposal. Until the last ‘FC’ (fully connected) layer, before ‘Softmax Activation’ layer, the feature vector of size 1-by-1000 dimension will be extracted. This process is performed in step (d) as shown in figure 1. The features of every salient region are then concatenated as one final feature vector of size 1-by-$(n \times 1000)$ dimension, where $n$ represents the number of salient regions.

3.3. Multi-class Linear Support Vector Machines Classification
The classification process of this model is done by Multi-class Linear Support Vector Machine (SVM) classifiers as indicated in figure 1 step (e). SVM is a very useful state-of-the-art supervised technique that commonly used for data classification problem [28]. For the proposed model in this study, the implementation version of multi-class SVMs introduced in [29] is being adopted. The multi-class classification is done with $r$ linear SVM classifiers, where $r$ is the number of scene categories.
categories involved. These linear SVMs are trained with one-vs-all method. Every binary linear SVM is solving the following unconstrained convex optimization problem

$$
\min_{w_c} \left\{ J(w_c) = \|w_c\|^2 + C \sum_{i=1}^{n} l(w_c; y_i^c, z_i) \right\}
$$

(1)

where $y_i^c = 1$ if $y_i = c$, else $y_i^c = -1$. $l(w_c; y_i^c, z_i)$ is a hinge loss function.

A differentiable quadratic hinge loss in equation (2),

$$
l(w_c; y_i^c, z_i) = \left[ \max(0, w_c^T z_i - y_i^c - 1) \right]^2
$$

(2)

is adopted since the standard hinge loss function is not differentiable everywhere. This also can slow down the use of gradient-based optimization methods. By adopting the differentiable quadratic hinge loss, the process of training can be done effortlessly with simple gradient-based optimization methods. The final deep feature vector (1-by-(n x 1000) dimension) generated from previous step is feed into these linear SVMs. The category of Linear SVM with highest confidence score will be assigned to the input scene image.

3.4. Model Training and Testing

The proposed model is trained and tested on MIT Indoor 67 scene image dataset [30]. This dataset is a very challenging dataset which has commonly been used to test the recognition performance on indoor scene. This dataset originally containing 67 scene categories with a total of 15,620 images. The number of images varies across categories. Figure 2 shows the example images of 15 different scene categories from MIT Indoor 67 database.

The process of generating salient region proposals of scene images is performed with image cropping process, hence larger number of scene images unfortunately produce experimental difficulties. Therefore, in this study, only 15 random categories with 100 images per category (80 training images, 20 testing images) will be used to train and test the performance of our proposed model.

All images will undergo the process of salient region generation by Saliency Toolbox as explained in section 3.1. $n$ number of salient regions cropped from its original image will be feed into the GoogLeNet CNN pre-trained model one by one to extract its deep features. The deep features of every salient region are concatenated as one final deep feature before it is used to train the multi-class classification with $r$ linear SVM classifiers indicating number of scene category involved.

The trained SVMs are tested with images of testing set. The category label of testing image is assigned to SVM with the highest confidence score. Therefore, the overall classification performance of this model is calculated by the average multi-class classification accuracy of all scene categories. Accuracy of every scene category is calculated using equation (3);

$$
\text{accuracy}, a = \frac{c}{t} \times 100
$$

(3)

where $c$ is number of correctly classified image(s) and $t$ is total number of images.

Hence, the overall average accuracy of the model is calculated using equation (4);

$$
aa = \frac{\sum_{i=0}^{n} a_i}{r}
$$

(4)

where $aa$ is average accuracy and $r$ is total number of scene category involved.
4. Experiments
In this section, an experiment and a benchmarking testing will be conducted to the proposed model. For the experiment, the proposed model will be tested to perform classification task under three different settings. Setting that achieved the best classification accuracy will be used as the default setting of the model. For the benchmarking testing, we evaluate the performance of our proposed model by comparing its average accuracy [2][31] to average accuracy achieved with standard normal approach model. Olson et.al in [32] explained that the term benchmarking is referring to the evaluation and comparison of the ability of different methods.

4.1. Experiment 1: Model’s Ability Evaluation
In this experiment, the performance of the proposed model will be tested with difference number of salient region proposals; (i) Setting 1: Three region proposals, (ii) Setting 2: Five region proposals, and (iii) Setting 3: Ten region proposals. Different number of region proposals used might affect the final average accuracy performance achieved by the model. It is expected that the more salient region proposals used, the better the classification performance of the model.

In this experiment, the images from MIT-Indoor 67 dataset has been used. The process of training and testing the model are repeated for every setting as below;
Setting 1: 3 salient region proposals
1. For every image of training and testing set, the first 3 salient regions generated by Saliency Toolbox will be used (refer to section 3.1 for the details of this process);
2. Every region is feed into GoogLeNet CNN model and 1000-dimensional vector of feature representation is extracted from last ‘FC’ (Fully Connected) layer. This step produced three deep feature vectors represent all the three regions (refer to section 3.2 for the details of this process);

![Figure 2. Example images of 15 different scene categories from MIT Indoor 67 database.](image-url)
3. The 1000-dimensional feature vectors of the regions are then concatenated as final 1-by-3000(1000x3) dimensional vector as the deep features to represent one scene image from the dataset;

4. These final 1-by-3000 feature vectors of every training image is used to train model’s multi-class linear SVM classifiers. The deep feature vectors of every testing image generated with the same procedure are used to validate the trained classifiers (refer to section 3.3 for the details of this process);

5. Accuracy of every scene category is calculated using equation (3). Hence, the overall average accuracy of the model is calculated using equation (4). These steps are repeated for Setting 2 (5 regions) and Setting 3 (10 regions). The accuracies obtained with different settings are compared and evaluated. The model will be finalized to operate with the setting that achieved highest accuracy.

4.1.1. Result
In the table 1, the average accuracy of the proposed model is increased as the number of salient regions used, increased. In Setting 3, by using ten salient region proposals, the model has achieved the best performance with 90.67% average accuracy. Setting 2 is 84.00% average accuracy, and Setting 1 with three regions as expected achieved the lowest average accuracy, 79.33%.

The result has clearly shown that the more salient regions used, the better the average accuracy performance of the model. More regions would provide more local discriminative information about scene image. This is further discussed in section 5.1.

Based on this results, the scene classification model will operate using Setting 3, ten salient region proposals. Benchmarking testing of this model in the following section will be fixed to use this setting.

Table 1. Accuracy percentage for every scene and average accuracy achieved by the model under three different settings.

| Scene          | Setting 1 (3 Salient Objects) | Setting 2 (5 Salient Objects) | Setting 3 (10 Salient Objects) |
|----------------|-------------------------------|-------------------------------|-------------------------------|
| Auditorium     | 75                            | 90                            | 100                           |
| Bakery         | 85                            | 90                            | 95                            |
| Bookstore      | 80                            | 95                            | 100                           |
| Buffet         | 70                            | 75                            | 85                            |
| Children room  | 85                            | 75                            | 70                            |
| Closet         | 80                            | 85                            | 95                            |
| Computer room  | 85                            | 95                            | 95                            |
| Florist        | 80                            | 85                            | 95                            |
| Greenhouse     | 85                            | 95                            | 95                            |
| Gym            | 80                            | 95                            | 95                            |
| Operating room | 80                            | 80                            | 85                            |
| Pantry         | 80                            | 70                            | 85                            |
| Prison cell    | 70                            | 75                            | 80                            |
| Music studio   | 100                           | 95                            | 100                           |
| Wine cellar    | 55                            | 60                            | 85                            |
| Average Accuracy | 79.33                        | 84.00                        | 90.67                        |

4.2. Experiment 2: Benchmarking Testing
For further evaluation of the model’s performance, benchmarking testing is conducted. The proposed model is set up to operate with setting 3, ten salient region proposals generated by Saliency Model. The performance the proposed model is evaluated by comparing its accuracy to the accuracy achieved
by standard normal approach model, where the Saliency Toolbox is not being used. This means, the setting for the standard normal approach scene classification model is based on entire whole image. The standard normal approach model is not performing classification based on salient objects, instead based on entire image.

With the same set of images from MIT Indoor 67 dataset, the standard model takes input images straight to GoogLeNet and generates its deep feature. A set of 1-by-1000 dimensional deep feature of every image is then extracted from the last ‘FC’ (Fully Connected) layer. Extracted features for training images are used to train the multi-class linear SVMs. The features for testing images generated using the same procedures are used to validate the trained classifiers. Accuracy of every scene category is calculated using equation (3). Hence, the overall average accuracy of the standard normal approach model is calculated using equation (4).

4.2.1. Result
As shown in table 2, the proposed model has achieved an average accuracy of 5.67% higher than the standard normal approach model which classifies just based on whole entire image. Standard normal scene classification model only achieved 85% average accuracy.

Table 3 shows some scene categories: Children room, Buffet and Operating room, are observed to have a decrease in accuracy using the proposed model compared to standard approach model. This might be due to some of the salient objects attended from the images are also attended from images of other categories. Therefore, the representation of some images of these categories are not strong enough to distinguish from other categories. Some categories: Bakery, Computer room and Prison cell have made a huge positive improvement compared to standard approach model. Besides these three categories, the other five categories have also made an improvement by at least 5%. These results have proven that classifying scene images based on salient regions would help improving accuracy performance. These findings will be further discussed in the next section.

Table 2. Average accuracy achieved by the proposed scene classification model and standard approach model.

| Approach                               | Average Accuracy, % |
|----------------------------------------|---------------------|
| Proposed Scene Classification Model    | 90.67               |
| Standard Normal Scene Classification Model | 85.00               |

Table 3. Accuracy different for several scene categories from standard approach to proposed model.

| State       | Scene            | Performance change |
|-------------|------------------|--------------------|
| Increase    | Bakery           | 75 → 95 (+20%)     |
|             | Computer Room    | 75 → 95 (+20%)     |
|             | Prison Cell      | 65 → 80 (+15%)     |
| No changes  | Closet           | 95                 |
|             | Green House      | 95                 |
| Decrease    | Children Room    | 75 → 70 (-5%)      |
|             | Buffet           | 95 → 90 (-5%)      |
|             | Operating Room   | 90 → 85 (-5%)      |

5. Discussion

5.1. Number of Region Proposals
The Experiment 1 indicated that the number of region proposals is significantly affecting the classification performance of the model. These results support the hypothesis that the more region
proposal used, the better the classification performance. Providing more region proposals means providing more information of an input image, this will assist the classifier to give good classification output. This idea follows the concept of human ability when performing classification task. As human classifies scene by observing objects present, the more objects human can see in an image, the more information they able to collect, hence at once lead their mind for an excellent decision making.

Since this study is specifically focusing on modelling scene classification which process based on human pre-attentive visual attention, the amount of salient proposals will be maximized to ten. In a glance, human only able to attend to several objects [6][7], means not all objects will be processed, yet still able to perform good classification output.

Furthermore, it is unnecessary to include all objects in scene image because only certain amount of objects are useful for the classification process [3], especially indoor images which contain a lot of different objects. Including inessential objects/regions is believe will increase the computational process complexity as the model will need to learn complex feature.

5.2. Global vs Local Feature Extraction
As explained in our model architecture, this approach is using pre-trained CNN model as a generic deep feature extractor. This approach is applied in some scene classification approaches since the original architecture of CNN model is not suitable for scene images and it is built specifically to process and classify only object images [5][20]. Features produce by the original CNN model on scene images are too heavy and disordered to represent the indoor scene [3][5]. Unlike some other approaches that believe scene recognition as straightforward as object recognition problem and applying original pre-trained CNN model mainly on global fine-tuning which take the entire whole scene images as input [1], our proposed model performs deep features extraction on local salient objects generated by Saliency Toolbox.

These findings show that local salient regions generated by Saliency Toolbox could provide meaningful and useful information of a scene image. Extracting from one whole image only produce noise features [3] and this does not keep much information of local feature distributions of the image [2]. Different scene environment is made up from different objects. The global feature representation cannot provide subtle information about the scene images and it can complicate the learning and classification process [20].

Scene image that contains a lot amount of different objects will provide broad and hazy features. It is too general to be used to train and test the classifiers. For many classification tasks, a strong deep feature of local objects that holds very precise and specific information about the scene are important. The objects are said to be the discriminative characteristic since the presence of objects in scene environment is different between categories. Besides, this approach can resolve the problem of global feature representation. Therefore, this study has demonstrated how the problem of semantic gap happened in scene classification can be reduced.

Apart from that, the well performance of the proposed model has been supported by Saliency Toolbox model which able to produce important regions holding meaningful distinctive information of salient object.

5.3. Saliency Toolbox as Region Proposal Generator
In this study, one of the earliest saliency model, Saliency Toolbox [26], has been used to generate the region proposal for an indoor scene image. The findings from Experiment 2 have suggested that Saliency Toolbox is capable to produce high-quality regions that likely containing salient object(s). The extracted object(s) bring meaningful local semantic information details about an input scene image. The results show that, in many different objects available in indoor scene, our proposed model is only using ten (10) salient objects could achieved an average accuracy up to 90.67%.

Some studies on scene classification problem have applied the approach of using objects representation to classify scene images [8][9][10][11][12]. The existing methods were trying to generate unlimited amount of objects available within a scene image, which clearly contradict the
concept of human visual attention and perception. These approaches are not specifically focusing on the idea of how human bottom-up visual attention is important for scene recognition. The Saliency Toolbox is proposed in study in order to include human visual information processing mechanism when generating region proposals and sub-sequentially use the regions to improve scene recognition problem.

5.4. Computational Complexity
The computational complexity of the proposed approach with different settings and standard normal approach is computed on 13-inch MacBook Air with the following specifications: 1.3 GHz Intel Core i5 processor, 4GB 1600 MHz DDR3 Memory, NVIDIA GeForce GTX 1050 Ti 4095MB eGPU, running macOS High Sierra Version 10.13.6 operating system. Both proposed and standard approach models is implemented in MATLAB R2018a. For feature extraction and concatenation process, the period of time in second is shown in table 4.

| Approach               | Time, s |
|------------------------|---------|
| Setting 1 (3 salient objects) | 38      |
| Setting 2 (5 salient objects) | 57      |
| Setting 3 (10 salient objects) | 100     |
| Standard Normal Model  | 30      |

6. Conclusion
A new approach for scene classification that built on top of pre-trained CNN model is proposed. This approach is unlike other scene classification approaches, the advantage of saliency model to generate salient region proposals is integrated into the classification model to improve the classification performance. The experiment conducted shows the proposed model generates better result. The Saliency Toolbox is able to provide meaningful regions/objects for scene classification. Nevertheless, this study also proved that providing more region proposals will lead to more accurate classification output.

Acknowledgement
This study was partially supported by Zamalah Unimas Postgraduate Scholarship (ZSU), (February 2017 – September 2018) provided by University Malaysia Sarawak and we would also like to thank Faculty of Cognitive Sciences and Human Development for supporting this work.

References
[1] Seong H, Hyun J and Kim E 2020 FOSNet: an end-to-end trainable deep neural network for scene recognition IEEE Access 8 82066-82077.
[2] Liu W, Li Y and Wu Q 2018 An attribute-based high-level image representation for scene classification IEEE Access 7 4629-4640.
[3] Li Y, Zhang J, Cheng Y, Huang K and Tan T 2018, April DF 2 Net: Discriminative feature learning and fusion network for RGB-D indoor scene classification Thirty-Second AAAI Conference on Artificial Intelligence.
[4] Wolfe J M and Horowitz T S 2017 Five factors that guide attention in visual search Nature Human Behaviour 1(3) 1-8.
[5] Guo W, Wu R, Chen Y and Zhu X 2018 Deep learning scene recognition method based on localization enhancement Sensors 18(10) 3376.
[6] Failing M and Theeuwes J 2018 Selection history: How reward modulates selectivity of visual attention Psychonomic Bulletin & Review 25(2) 514-538.
[7] Wang W and Shen J 2017 Deep visual attention prediction IEEE Transactions on Image Processing 27(5) 2368-2378.
[8] Wu R, Wang B, Wang W and Yu Y 2015 Harvesting discriminative meta objects with deep CNN features for scene classification Proceedings of the IEEE International Conference on Computer Vision 1287-1295.
[9] Zhao Z and Larson M 2018, October From volcano to toyshop: Adaptive discriminative region discovery for scene recognition Proceedings of the 26th ACM International Conference on Multimedia 1760-1768.
[10] Herranz L, Jiang S and Li X 2016 Scene recognition with CNNs: objects, scales and dataset bias Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 571-579.
[11] Cheng X, Lu J, Feng J, Yuan B and Zhou J 2018 Scene recognition with objectness Pattern Recognition 74 474-487.
[12] Sun N, Li W, Liu J, Han G and Wu C 2018 Fusing object semantics and deep appearance features for scene recognition IEEE Transactions on Circuits and Systems for Video Technology 29(6) 1715-1728.
[13] Redmon J and Farhadi A 2017 YOLO9000: better, faster, stronger Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 7263-7271.
[14] Jiafa M, Weifeng W, Yahong H and Weiguo S 2019 A scene recognition algorithm based on deep residual network Systems Science & Control Engineering 7(1) 243-251.
[15] Oliva, A and Torralba A 2001 Modeling the shape of the scene: A holistic representation of the spatial envelope International Journal of Computer Vision 42(3) 145-175.
[16] Lowe D G 2004 Distinctive image features from scale-invariant keypoints International Journal of Computer Vision 60(2) 91-110.
[17] Pandey M and Lazebnik S 2011, November Scene recognition and weakly supervised object localization with deformable part-based models. IEEE 2011 International Conference on Computer Vision 1307-1314.
[18] Juneja M, Vedaldi A, Jawahar C V and Zisserman A 2013 Blocks that shout: Distinctive parts for scene classification Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 923-930.
[19] LeCun Y, Bengio Y and Hinton G 2015 Deep learning Nature 521(7553) 436-444.
[20] Khan S H, Hayat M, Bennamoun M, Togneri R and Sohel F A 2016 A discriminative representation of convolutional features for indoor scene recognition IEEE Transactions on Image Processing 25(7) 3372-3383.
[21] Li N, Zhao X, Yang Y and Zou X 2016 Objects classification by learning-based visual saliency model and convolutional neural network Computational Intelligence and Neuroscience 2016.
[22] Wang X, Zhong Y, Xu Y, Zhang L and Xu Y 2018 Saliency-based endmember detection for hyperspectral imagery IEEE Transactions on Geoscience and Remote Sensing 56(7) 3667-3680.
[23] Zanca D and Gori M 2017 Variational laws of visual attention for dynamic scenes Advances in Neural Information Processing Systems 3823-3832.
[24] Almeida A F, Figueiredo R, Bernardino A and Santos-Victor J 2017, November Deep networks for human visual attention: A hybrid model using foveal vision Iberian Robotics Conference (Springer, Cham) 117-128.
[25] Navalpakkam V, Arbib M and Itti L 2005 Attention and scene understanding Neurobiology of Attention (Academic Press) 197-203.
[26] Walther D and Koch C 2006 Modeling attention to salient proto-objects Neural Networks 19(9) 1395-1407.
[27] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A 2015 Going deeper with convolutions Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 1-9.
[28] Mehmood Z, Mahmood T and Javid M A 2018 Content-based image retrieval and semantic
automatic image annotation based on the weighted average of triangular histograms using support vector machine Applied Intelligence 48(1) 166-181.

[29] Yang J, Yu K, Gong Y and Huang T 2009, June Linear spatial pyramid matching using sparse coding for image classification 2009 IEEE Conference on Computer Vision and Pattern Recognition 1794-1801.

[30] Quattoni A and Torralba A 2009, June Recognizing indoor scenes 2009 IEEE Conference on Computer Vision and Pattern Recognition 413-420.

[31] Herranz L, Jiang S and Li X 2016 Scene recognition with CNNs: objects, scales and dataset bias Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 571-579.

[32] Olson R S, La Cava W, Orzechowski P, Urbanowicz R J and Moore J H 2017 PMLB: a large benchmark suite for machine learning evaluation and comparison BioData Mining 10(1) 1-13.