Journal of Positive School Psychology http://journalppw.com
2022, Vol. 6, No. 12, 2207 – 2217

# Scene Recognition using Multi-Label Multi-Resolution CNN

**Mehrdad Jannesar**

*Master of artificial intelligence, islamic azad university of south tehran branch, Iran/Tehran*

## Abstract

Classification is one of the challenging issues of machine Learning and it is considered as one of the most complex issues in this area. In scene classification, feature extraction is one of the most important parts of every algorithm; because extraction of appropriate features will increase the accuracy of classification. Regarding the ambiguity and variety of the contents of images which have a poor appearance due to the changes of light and scale, scene recognition is more challenging than object recognition. In this paper, a new scene recognition method is proposed based on multi-label multi-resolution- CNN. Our method is evaluated by both indoor and outdoor scene datasets. The superior performance of our method is proved by simulation results and compared with other state of arts.

**Keywords:** Scene recognition, Machine Learning, CNN, Feature extraction

## I Introduction

Scene classification is a major problem in computer vision and researchers have proposed many solutions in the past few years [1-3]. One of the main mysteries of vision is the significant capacity of human brain in understanding the new scenes, places, and events in a quick and effortless manner. Regarding the ambiguity and variety of the contents of images which have a poor appearance due to the changes of light and scale, scene recognition is more challenging than object recognition.

Image classification has been the center of image processing studying for many years. Various practical methods have been proposed in different problems. Traditional machine learning algorithms such as Bayesian [4], support vector machine (SVM) [5], [6], random forest (RF) [7] are used to classify an image based on the features extracted which have some limitations [8]. Recently, deep convolutional neural network (CNN) has a widespread use for image classification, which makes great progress in many different fields [9-10].

In general, scene recognition consists of both semantic and contextual information that are used in other image processing issues such as object detection [11] and action classification [12]. Scene classification considered to be challenging due to inherent uncertainty of scene concepts and increasing similarities among different categories. In other words, some class of scenes have a similar visual appearance and are easily misclassified with some other labels. This will become worse as the number of classes in the scene increases. Accordingly, in this research we propose a novel Multi-label deep reinforcement learning for scene classification method which consist of two major modules: CNN visual attention which is inspired by the state of art in [13] and label view by using large margin nearest neighbor CNN (LMNN-CNN). In this project we try to overcome scene recognition problem in large scale scene database which is mentioned in [14].

Scene recognition in a set of large-scale data in which include hundreds of classes and millions of images, has many challenges because of the nature uncertainty and the high overlapping between different classes. Visual inconsistence and label ambiguity are the important challenges in the data set [14].

**Visual inconsistence**: this challenge means that for a group with same scene (for example

kitchen) there is a wide range of images. This rises a problem to detect an image group as an objective.

**Label ambiguity**: some of the scenes group have the similar visualization and can be easily replaced with some other groups. Fig.1. shows some of these mistakes. This error in the scene understanding could be a bigger problem when the number of scene classes increases. In other word, in these situations, the inter-class overlapping will be problematic.


Fig. 1 Example of incorrect labeling in Places 401 dataset using method in [14].

According to mentioned problems, in this paper, a method based on the multi-label learning and multi-resolution CNN has been presented to overcome the wrong labeling problem in the scene classification. In our method we consider both label-view (low dimensional) features by using LMNN-CNN and feature-view (deep features using multi-resolution CNN) in order to obtain all infrastructures information of target scenes. The rest of paper is organized as follows:

In Section 2 we quickly explain the multi-resolution CNN structure in scene classification and then propose our enhanced approach. In Section 3 we evaluate our propose method and compare it with other state of arts in [3], [16], [20] and [23-25]. Finally, in Section 4 the conclusion is represented.

## 2 Multi-resolution CNN

Generally, a visual scene can be considered as a combination of objects and semantic levels along each other in an understandable form. In other word, the scenes have meaningful elements which are visible in the different middle layers. For this purpose, the weak- and strong-level features' properties with different image resolution were used to create the model of the present paper. Fig. 2 shows the structure of a multi-resolution network. As can be seen from this figure, the network is composed of a double-layer convolution and max pooling layers. This structure includes two small-scale and large-scale modules in which the batch normalization has been used [15].

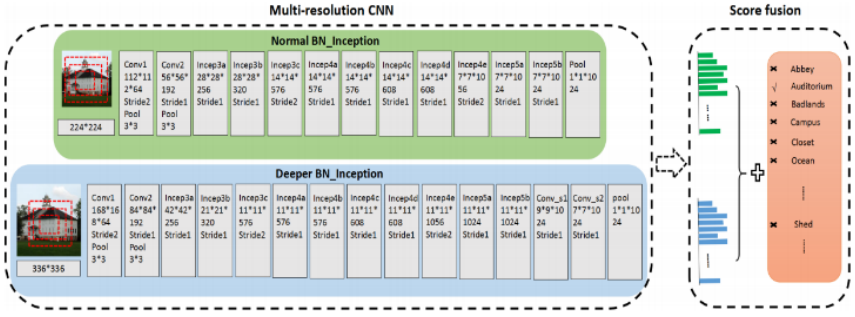


Fig. 2 The structure of the multi-resolution CNN which includes two low- and high-resolution CNN modules.The normal module extracts the image structure with large-scale and the deep module

describes the visual patterns at small scales and finally performs the final recognition by combining the rating as the averaging [14].

**Large-scale CNN module:** this module works on images with 224*224 dimensions and includes 3 layers with pre-trained weights. The structure of this module is similar to the module presented in [15]. The main focus of this module is on the describing the overall arrangement of the objects in the image.

**Small-scale CNN module:** in this module, for the images with resolution higher than 384*384, the process performs on regions with size 336*336. With considering the larger images as input, the network layers' deep can be increased which increases the model robustness. In this module, as can be seen from Fig. 2, three convolution layers have been added compare with large-scale module. In this module, the padding has not been done too and the final feature space size (before global average pooling) is 7*7. This module is responsible for providing the image information as the micro structures and enables us to access more accurate local details.

The two modules mentioned above received the images with different resolution as an input and their corresponding layers have layers with different sizes that eventually by averaging the predicted ratings from both modules the decision can be made on the image label.

The descriptions provided are for double-resolutions CNN network, where the images are trained using two resolutions of 256*256 and 336*336. This idea can be extended to multi-resolution case and can expects to have better recognition result. It should be noted that in the described multi-resolution CNNs learning, each module is trained independently. The common steps used in [8] and [16] have been used for training which will be described in Section 3 with more details.

### 2.1 Proposed Multi-label Multi-resolution CNN

Reviewing the work [14] shows that there is still ambiguity problem in labeling some of the images which some of them are presented in Fig. 1. In this paper, the multi-resolution CNNs network model combined with multi-label learning is proposed to overcome this problem.

Multi-labeling learning plays an important role in the machine learning [17] and data mining [18] fields. Unlike the common classifying problems, in multi-label training each sample is related to some labels simultaneously. In the real-world problems, the number of possible labels of a sample can be too many for a common group and achieving a very useful labeled data set for training the model is hard.

In the field of multi-label object recognition, different levels of labeling are defined. In other words, weak labels only define the existence of the object while the strong labels include the main components of the object in the reference image. An important fact that should be paid attention is that the weak labels can be used for accurate tuning of a pre-trained CNN network in order to produce good usable global representatives.

In this paper, the weak and strong labels were used at the same time. For this purpose, the weak labels are used for accurate tuning of a multi-resolution network and extract the global features which can improve the classification performance along with strong labels as view labels.

The general framework of multi-label multi resolution CNNs presented in this paper has been shown in Fig. 3. In order to describe the recommended framework, it is assumed that by having $n$ learning images as $\{\boldsymbol{X}_i\}_{i=1}^{n}$ we can extract $n_i$ proposals as $\{\boldsymbol{x}_{ij}, j = 1, \dots, n_i\}$ from each of the $\boldsymbol{X}_i$ images using the unsupervised selective search method [19].

Then these images were used in order multi-label training of a multi-resolution CNN network with two resolutions 256 and 385.

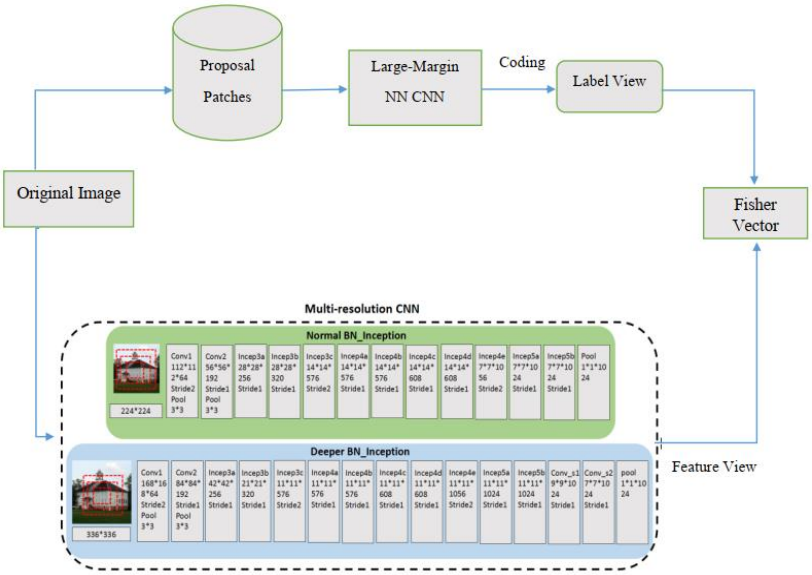


Fig. 3 Schematic of proposed multi-label multi resolution CNN framework. First, by extracting the proposals from images with two different resolutions, the image recognition problem is defined as multi-classes-multi-samples training problem and then two feature classes are extracted. The first feature are the features with small dimensions of the LMNN-CNN which produces the label view features by barcoding the k[th] label information of the nearest neighbor from the image pool which includes ground truth images. The second category includes a double-resolution CNN network which is responsible for producing the conceptual features. Finally, the final feature vector is created by combining these two features using the Fisher vector.

Also, in order to extract the label view features a CNN network of a LMNN-CNN was used. In this network, the cost function logistic is replaced with the cost function of the nearest neighbor using the following relation [20].

$$\sum_{i,j} \eta_{ij} D(x_i, x_j) + C \sum_{i,j,l} \eta_{ij} (1 - y_{il}) [1 + D(x_i, x_j) - D(x_i, x_l)]_+$$

Where y is the label information such that if $x_i$ and $x_k$ belong to the same class, then $y_{ik} = 1$, otherwise $y_{ik} = 0$. Parameter C is tradeoff parameter and $[.]_+ = \max(\cdot, 0)$ is the hinge loss function. The learning distance metric is $D(x_i, x_j) = \|W(x_i - x_j)\|^2$. And η is also considered as the following function.

$$\eta_{ij} = \begin{cases} 1 & \text{if } x_j \text{ is one of the } k \text{ nearst neighbor of} x_i \\ 0 & otherwise \end{cases}$$

The output of the LMNN-CNN is as the features with small dimension which has both the CNN semantic features and the LMNN good neighboring features. The details of the used parameter related to the training the networks are described in the next chapter.

Finally, the labeling view and feature view features are combined using the Fisher vector defined in [20] and the final feature vector is obtained. The Gaussian mixture model (GMM) was used to create the Fisher vector in order to classifying the images. It is assumed that the GMM model with K elements has $\lambda = \{\omega_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, k = 1, \dots, K\}$ where $\omega_k$, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are mixture weight, mean vector and covariance matrix of the k[th] Gaussian model. By assuming the diagonal covariance matrix with diagonal elements $\sigma_k$ for forming the Fisher vector (FV), we have

$$\boldsymbol{f}_{\omega_k}^{\boldsymbol{X}_i}$$

$$= \frac{1}{\sqrt{\omega_k}} \sum_{j=1}^{n_i} \left(\gamma_j(k) - \omega_k\right)$$

$$\boldsymbol{f}_{\boldsymbol{\mu}_k}^{\boldsymbol{X}_i} = \frac{1}{\sqrt{\omega_k}} \sum_{j=1}^{n_i} \gamma_j(k) \left(\frac{\boldsymbol{x}_{i,j} - \boldsymbol{\mu}_k}{\sigma_k}\right)$$

$$\boldsymbol{f}_{\sigma_k}^{\boldsymbol{X}_i} = \frac{1}{\sqrt{\omega_k}} \sum_{j=1}^{n_i} \gamma_j(k) \frac{1}{\sqrt{2}} \left(\frac{(\boldsymbol{x}_{i,j} - \boldsymbol{\mu}_k)^2}{{\sigma_k}^2} - 1\right)$$

where, $\gamma_j(k)$ is the soft weight. By using the all above relation, all proposals $\{\boldsymbol{x}_{ij}, j = 1, \dots, n_i\}$ of the image $\boldsymbol{X}_i$ are mapped in to FV. Finally, by putting the feature view and label view together as $[\boldsymbol{f}_{ij} \ \lambda \boldsymbol{l}_{ij}]$ for every proposal $\{\boldsymbol{x}_{ij}, j = 1, \dots, n_i\}$, we reach the final vector's form. The parameter $\lambda$ considered as a trade-off between the two categories.

## 3 Simulation Results

### 3.1 Image dataset

Two benchmark datasets of MIT Indoor-67 and VOC 2007 have been used in our work.

- **MIT Indoor-67 dataset**

This dataset includes 15620 images of 67 different indoor spaces. This dataset has a lot of challenges in separating different classes such as conference rooms and theater halls. In this paper, as [21], we have used 80 images from each class for training and 20 images for testing.

**VOC 2007 dataset**

This dataset includes 9963 images of 20 classes in different environments [22].

### 3.2 The results of simulation

As stated before, in this paper, the pre-trained two-resolution CNN has been used with the label features obtained from LMNN-CNN (as [20]). The properties of the used two-resolution CNN network are presented in Figure 2. As seen in figure 2, a network includes the images with a resolution 224*224 and the other network includes the images with a resolution of 336*336. In this work, stochastic gradient decent (SGD) algorithm has been used for training the two-resolution CNN network. The parameters used for SGD algorithm are also presented in Table 1.

Table 1. The parameters used in SGD algorithm

| Initial learning rate | Learning rate drop factor | Learning rate drop factor per Epochs | Number of Epochs | Momentum |
|---|---|---|---|---|
| 0.001 | 0.1 | 4 | 20 | 0.9 |

The next part includes the label features and as explained in the previous chapter, they are resulted from LM-CNN network, and they are coded in a binary manner. The structure of LM-CNN network is presented in Table 2. The stages of coding the neighboring labels are the same as [20]. As stated in the previous section, having *n* training images as $\{\boldsymbol{X}_i\}_{i=1}^{n}$, we extract $n_i$ sample objects or proposals $\{\boldsymbol{x}_{ij}, j = 1, \dots, n_i\}$ of each of the $\boldsymbol{X}_i$images by using unsupervised selective search method [19]. Figure. 4 presents some examples of the sample objects extracted by using unsupervised selective search algorithm for two images of VOC dataset.

Table 2. The structure of large margin nearest neighbor CNN network

| Conv1 | Conv2 | Conv3 | Conv4 | Conv5 | Full6 | Full7 |
|---|---|---|---|---|---|---|
| 96*7*7<br>Stride 2<br>Padding 0<br>*2 pool | 256*5*5<br>Stride 2<br>Padding 1<br>*2 pool | 512*3*3<br>Stride 1<br>Padding 1 | 512*3*3<br>Stride 1<br>Padding 1 | 512*3*3<br>Stride 1<br>Padding 1<br>*2 pool | 4096<br>dropout | 128<br>dropout |

Fig. 3. The selected parts as the proposal images

Regarding the mentioned structures of CNN networks, it is obvious that the high dimensions of feature vectors can make the classification difficult. Therefore, principle component analysis (PCA) algorithm is used in different simulation scenarios.

Finally, the final label of the image is determined by using support vector machine classifier (SVM) with linear and RBF kernels. In this paper, we use the Python 3.7and its relative library.

The quantitative criteria used for evaluating the results of simulation include classification accuracy, false positive rate, and specificity. Classification accuracy is obtained by calculating the confusion matrix. Confusion matrix is a square matrix of C×C in which, C is the number of classes. The main diameter of this matrix includes the samples classified accurately and non-diameter elements represent for the samples inaccurately classified in a class other than their real class. After calculating the confusion matrix, classification accuracy is calculated as the mean of the diameter elements. Two further quantitative criteria are calculated as the following relations.

$$FPR = FalsePositiveRate = \frac{FP}{FP + TN}$$

$$Specificity = \frac{TN}{FP + TN}$$

We consider two different scenarios for evaluating our method. In the first scenario, the results of scene classification are presented regarding only the two-resolution CNN network. In this scenario, the results are investigated by using PCA dimensionality reduction and without using that. In the second scenario, the label features are included beside the features of multi-resolution network as [$\boldsymbol{fij}$ $\boldsymbol{\lambda lij}$] ($\lambda$=0.5 and the total dimensions of the feature vector are equal to 11883) and help to increase of accuracy. Furthermore, all the simulations have been done in Windows by using hardware including 24 GB DDR4 RAM, CPU Intel Core i7-6700K, and VGA GTX 1080.

### 3.3 Scene Classification using Two-resolution CNN

As stated before, in this section, two CNN networks with different dimensions of input images have been used and the higher rate is determined as the image label. The results of scene classification for training, testing, and comparing them with the works [3], [16], [20] and [23-25] are presented in Table 3. As seen in

the Table 3, 81% accuracy has been resulted in Indoor-67 dataset which is more than the accuracy of [24]. As it is observed, in MIT Indoor-67, regarding the high number of classes and more challenges of images, labeling confusion is so high. In VOC 2007 dataset, the highest accuracy is about 90% which is better than the accuracy of [16] and [25]; although [6] has had a higher accuracy. Figure 4 and Figure 5 respectively present confusion matrixes for VOC 2007 and Indoor-67 datasets.

Table 3. The results of quantitative criteria for the state of using two-resolution CNN

| methods | dataset | Accuracy |
|---|---|---|
| Two-resolution CNN | MIT Indoor-67 | 80.3 |
| Two-resolution CNN+PCA | | 81 |
| VSAD+FV+VGGNet16 [23] | | **86.2** |
| LS-DHM [3] | | 85.3 |
| Semantic FV [24] | | 72.9 |
| Two-resolution CNN | VOC 2007 | 89.6 |
| Two-resolution CNN+PCA | | 90 |
| Fusion [20] | | **92** |
| HCP-2000C [16] | | 85.2 |
| VeryDeep [25] | | 89.3 |

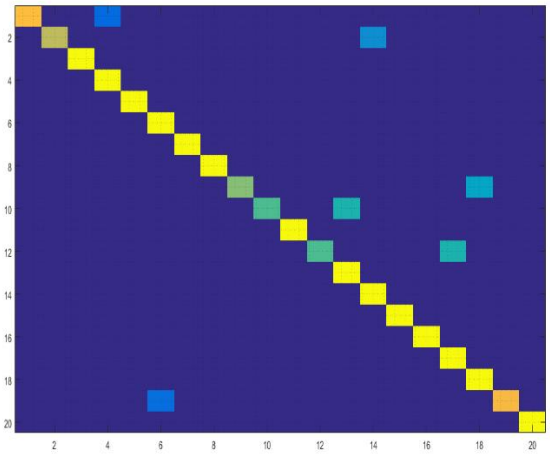

Fig. 4 Confusion matrix for a 31-class dataset with the accuracy of 90%

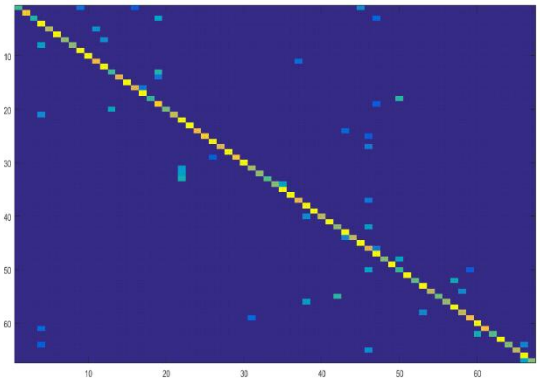

Fig. 5 Confusion matrix for a 67-class dataset with the accuracy of 81%

### 3.4 Scene Recognition by using the Proposed Method

This section provides the results of the proposed method of multi-label two-resolution CNN that was described in third chapter. As stated before, the feature vector is composed of the combination of the features of CNN connected layer and the (binary) label features resulted from LMNN CNN. Classification has been done by using support vector machine with linear kernels and radial basis function. The results of scene classification for testing and comparing them with the works [3], [16], [20] and [23-25] are presented in Table 4. As seen in the Table 4, in Indoor-67 dataset, the highest accuracy has been 86.4% that is higher than the accuracy of all works. In VOC 2007 dataset, a better performance than the other works was observed with the accuracy of 92%.

Table 4. The results of quantitative criteria for using the proposed method of multi-label two-resolution CNN

| methods | dataset | Accuracy | FPR | Specificity |
|---|---|---|---|---|
| multi-label two-resolution CNN | MIT Indoor-67 | 85.3 | 0.0022 | 0.9978 |
| multi-label two-resolution CNN+PCA | | **86.4** | 0.0021 | 0.9979 |
| VSAD+FV+VGGNet16 [23] | | 86.2 | ---- | ---- |
| LS-DHM [3] | | 85.3 | ---- | ---- |
| Semantic FV [24] | | 72.9 | ---- | ---- |
| multi-label two-resolution CNN | VOC 2007 | 91.9 | 0.0043 | 0.9957 |
| multi-label two-resolution CNN+PCA | | **92.3** | 0.0040 | 0.9960 |
| Fusion [20] | | 92 | ---- | ----- |
| HCP-2000C [16] | | 85.2 | ---- | ----- |

| | | | | |
|---|---|---|---|---|
| VeryDeep [25] | | 89.3 | ---- | ---- |

Figure 6 and Figure 7 respectively present the confusion matrixes for VOC 2007 and Indoor-67 datasets. As seen in figure 4-8, classes such as CHAIR, SOFA, and PLANT involve more challenges in classification; it is mainly due to the high similarity of the general properties of images of these classes with other classes. Also, in Indoor-67 dataset, classes such as concert hall, dental office, movie theatre, and meeting room involve more challenges in classification.

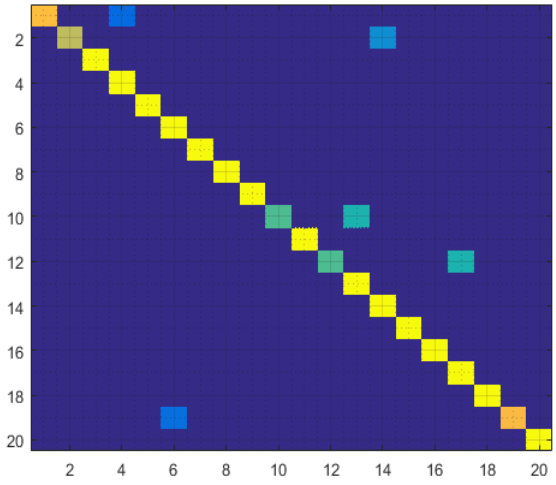


Fig. 6 Confusion matrix for a 20-class dataset with the accuracy of 92%

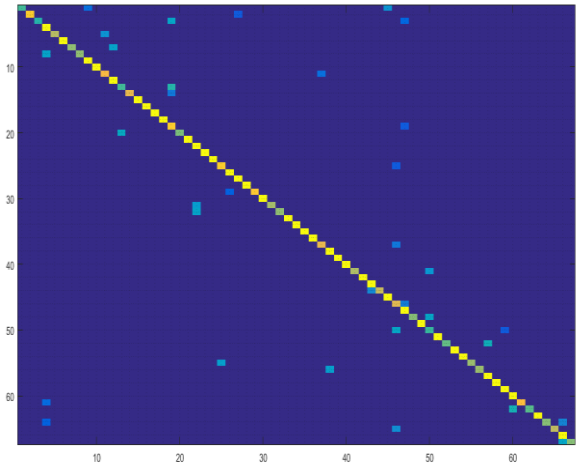


Fig. 7 Confusion matrix for a 67-class dataset with the accuracy of 86%

## 4 Conclusions

One of the main mysteries of vision is the significant capacity of human brain in understanding the scenes in a quick and effortless manner. In this paper, we presented a new multi-label multi-resolution framework to overcome mentioned image classification problem. Based on the proposed method on two bench mark image datasets (i.e. MIT-67 and VOC 2007) can be successfully transferred to tackle the multi-label problem. We evaluated our method on MIT-67 and VOC 2007, and verified that significant improvement can be made by multi-label multi-resolution CNN

compared with the state-of-the-arts. Furthermore, it is proved that dimension reduction can enhance the classification performance.

## REFERENCE


[1] L. Zhang, X. Zhen, and L. Shao, "Learning object-to-class kernels for scene classification," IEEE Trans. Image Processing, vol. 23, no. 8, pp. 3241–3253, 2014.

[2] Z. Zuo, B. Shuai, G. Wang, X. Liu, X. Wang, B. Wang, and Y. Chen, "Learning contextual dependence with convolutional hierarchical recurrent neural networks," IEEE Trans. Image Processing, vol. 25, no. 7, pp. 2983–2996, 2016.

[3] S. Guo, W. Huang, L. Wang, and Y. Qiao, "Locally supervised deep hybrid model for scene recognition," IEEE Trans. Image Processing, vol. 26, no. 2, pp. 808–820, 2017.

[4] D. Preotiuc-Pietro and F. Hristea, "Unsupervised word sense disambiguation with n-gram features," Artificial Intelligence Review, vol. 41, no. 2, pp. 241–260, 2014.

[5] O. Amayri and N. Bouguila, "A study of spam filtering using support vector machines," Artificial Intelligence Review, vol. 34, no. 1, pp. 73–108, 2010.

[6] B. Gu, V. S. Sheng, Z. Wang, D. Ho, S. Osman, and S. Li, "Incremental learning for -support vector regression," Neural Networks the Official Journal of the International Neural Network Society, vol. 67, no. C, pp. 140–150, 2015.

[7] Liaw and M.Wiener, "Classification and regression by randomforest," R news, vol. 2, no. 3, pp. 18–22, 2002.

[8] Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," arXiv preprint arXiv:1512.03385, 2015.

[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "Imagenet large-scale visual recognition challenge," International Journal of Computer Vision, vol. 115, no. 3, pp. 211–252, 2015.

[11] L. Wang, Y. Wu, T. Lu, and K. Chen, "Multiclass object detection by combining local appearances and context," in ACM Multimedia, 2011, pp. 1161–1164.

[12] Y. Zhang, L. Cheng, J. Wu, J. Cai, M. N. Do, and J. Lu, "Action recognition in still images with minimum annotation efforts," IEEE Trans. Image Processing, vol. 25, no. 11, pp. 5479–5490, 2016.

[13] Zhao, Dongbin, Yaran Chen, and Le Lv. "Deep reinforcement learning with visual attention for vehicle classification." *IEEE Transactions on Cognitive and Developmental Systems* 9, no. 4 (2016): 356-367.

[14] Wang, Limin, Sheng Guo, Weilin Huang, Yuanjun Xiong, and Yu Qiao. "Knowledge guided disambiguation for large-scale scene classification with multi-resolution CNNs." *IEEE Transactions on Image Processing* 26, no. 4 (2017): 2055-2068.

[15] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in ICML, 2015, pp. 448–456.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," CoRR, vol. abs/1409.1556, 2014.

[17] Yu, H. F., Jain, P., Kar, P., & Dhillon, I. (2014, January). Large-scale multi-label learning with missing labels. In *International conference on machine learning* (pp. 593-601).

[18] Kong, X., Ng, M. K., & Zhou, Z. H. (2013). Transductive multilabel learning via label set propagation. *IEEE Transactions on Knowledge and Data Engineering*, *25*(3), 704-719.

[19] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A.W. M. Smeulders. Selective search for object recognition. IJCV, 104(2):154–171, 2013.

[20] Yang, H., Zhou, J. T., Zhang, Y., Gao, B., Wu, J., & Cai, J. (2015). Can partial strong labwls boost multi-label object recognition?. *arXiv preprint arXiv:1504.05843*.

[21] Quattoni, Ariadna, and Antonio Torralba. "Recognizing indoor scenes." In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 413-420. IEEE, 2009.

[22] Everingham, Mark, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. “The pascal visual object classes (voc) challenge.” International journal of computer vision 88, no. 2 (2010): 303-338.
[23] Wang, Zhe, Limin Wang, Yali Wang, Bowen Zhang, and Yu Qiao. "Weakly supervised patchnets: Describing and aggregating local patches for scene recognition." IEEE Transactions on Image Processing 26, no. 4 (2017): 2028-2041.
[24] Dixit, Mandar, Si Chen, Dashan Gao, Nikhil Rasiwasia, and Nuno Vasconcelos. "Scene classification with semantic fisher vectors." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2974-2983. 2015.
[25] Wei, Yunchao, Wei Xia, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. "CNN: single-label to multi-label." arXiv preprint arXiv:1406.5726 (2014).