1) Scene Classification with Deep Convolutional Neural Networks

The paper [12] has a goal to demostrate the CNNs can help with the feature representation to extract high-level information of an image scene and thus improve the scene classification precision. They choose a CNN which is pre-trained on ImageNet dataset (ImageNet-CNN) since it is a large-scale general object recognition dataset. To utilize a pre-trained ImageNet CNN and for the efficiency of the feature extraction process, we use a popular library Caffe. For the training process, the system takes all images in the training set for each category as the input, use the ImageNet-CNN to perform a prediction for each image. Rather than getting the final 1000 length class prediction vector, they extract the response of Fully Connected Layer (FC) 7 of the CNNs, which is a 4096-dimensional vector contains 4096 response values.

An input image passes via ImageNet-CNN. The 4096 length deep feature vectors of ImageNet-CNN are very useful for the scene classification. They allot the one with most elevated certainty score.

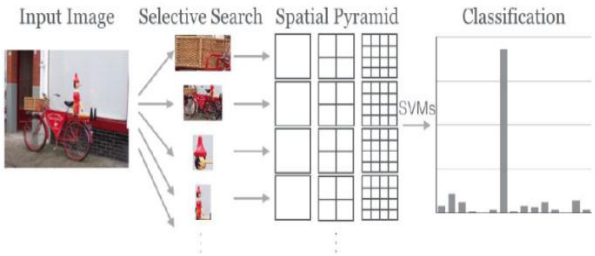


Figure 4: The overview of system [12]

As shown in above Fig. 4, a selective search algorithm is applied first to get roughly 2000 regions of interest for an input image. Then they apply a pre-trained Convolutional Neural Network (CNN) on each region of interest to get a deep feature vector of length. A three-level spatial pyramid representation of the image with deep features is used to create the final feature representation. At each level, for each spatial bin, we use max pooling to get the largest feature value of all the feature values of the regions of interest which fall into that spatial bin, resulting in the final feature of length 4096*(1+4+16) as a high-level representation of the input image. Then multiple one-vs-all linear SVMs are used to do the scene classification.

2) Basic level categorization facilitates visual object recognition

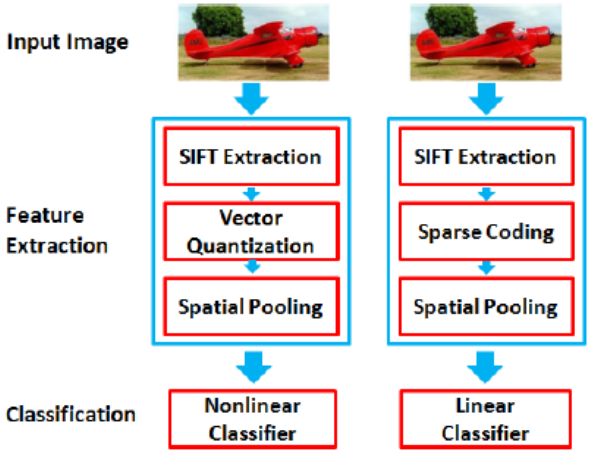
There has not been much research focused on linking CNNs with guiding principles from the human visual cortex. The paper [13] had idea of network optimization strategy inspired by both of the developmental trajectory of children's visual object recognition capabilities and Bar (2003), who hypothesized that basic level information is carried in the fast magnocellular pathway through the prefrontal cortex (PFC) and then projected back to inferior temporal cortex (IT), where subordinate level categorization is achieved. We instantiate this idea by training a deep CNN to perform basic level object categorization first, and then train it on subordinate level categorization. We apply this idea to training AlexNet [14] on the ILSVRC 2012 dataset and show that the top-5 accuracy increases from 80:13% to 82:14%, demonstrating the effectiveness of the method. We also show that subsequent transfer learning on smaller datasets gives superior results.

3) Support Vector Machine

In 1930s, Fisher proposed a procedure for theory of linear discriminants. The work of Frank Rosenblatt introduced perceptron learning rule in 1956 which give the advancement in fielf of artificial intelligence. In 1980s. Multilayer Perceptrons (MLP) are introduced which is very slow in learning. The SVM approach from is compared with the Linear discriminant Analysis (LDA) system and the results favor the SVM model with higher percentage of accuracy. The improved techniques were developed over the year, such as generative methods in [15-18] for modeling the co-occurrence of the codewords or descriptors, discriminative codebook learning in [19-22] instead of standard unsupervised K-means clustering, and spatial pyramid matching kernel (SPM) [23] for modeling the spatial layout of the local features.

4) Linear SPM based on sparse coding (ScSPM)

Author yang ad al. [24] proposed linear SPM based on sparse coding (ScSPM) then traditional nonliearn SPM method which is shown in Fig. 5.



(a) Nonlinear SPM (b) ScSPM

Figure 5: Schematic comparison of nonlinear SPM with ScSPM [24]

Any image is represented using set of descriptors. Using histogram and other statistics of descriptors, it is easy to compute a single feature vector.

In SPM approach, Z is a concatenation of local histograms in various partitions of different scales. Let zi denote the histogram representation for image Ii. For a binary image classification problem, an SVM aims to learn a decision function from eq. 1 .

$$f(z) = \sum_{i=1}^{n} \alpha_i k(z, z_i) + b \qquad (1)$$

where $\{(z_i, y_i)\}_i^n = 1$ is the training set. For a test image represented by z, the image is positively classified if function has positive value.

The approach of using linear SVMs based on SIFT was proposed by this paper. The U is result of applying the sparse coding to a descriptor set X and V is codebook to be pre-learned and fixed. The eq. 2 is use for calculating image feature.

$$z = F(U) \qquad (2)$$

Where the pooling function F is defined on each column of U. The pooling function F as a max pooling function on the absolute sparse codes by following Eq. 3.

$$Z_J = \max\{|u_{1j}|, |u_{2j}|, \ldots, |u_{Mj}|\} \qquad (3)$$

Where Zj is the j-th element of Z, u_{ij} is the matrix element at i-th row and j-th column of U. The number of local descriptors is represented by M. This max pooling outperforms other alternative pooling methods.

5) ResFeats: Residual Network Based Features for Image Classification

This paper [25] proposes new image features which are known as ResFeats which is shown in Fig. 6. This feature is extracted from the last convolutional layer of deep residual networks pre-trained on ImageNet. The applicability of it is in various image applications namely, object classification, scene classification. The result shows that ResFeats give better accuracy than their CNN counterparts on these classification tasks consistently. Because of the ResFeats are large feature vectors, this paper propose to use PCA for dimensionality reduction. Experimental results are provided to show the effectiveness of ResFeats with state-of-the-art classification accuracies on MLC datasets, Caltech-101 and Caltech-256.

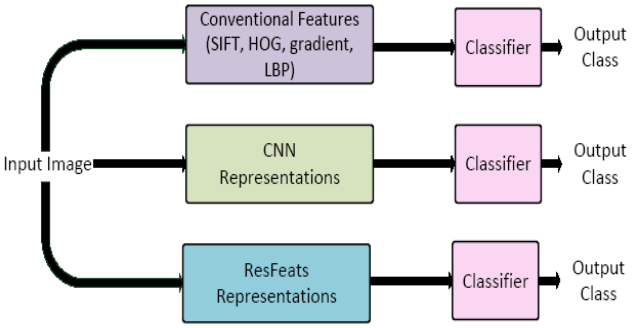


Figure 6: Evolution of classification pipelines [25]

Off-the-shelf ResFeats have the potential to replace the previous classification pipelines and improve performance for image classification tasks.

6) Local Coding Based Matching Kernel Method

The Bag of Visual Word (BoV) techniques are fundamentally simple. But, the main disadvantage is lower effectiveness. In opposition to it, kernel based metrics are more effective. But it has disadvantage of greater computational complexity and large memory requirements. This paper [26] demonstrates that a unified visual matching framework can be developing to include BoV and kernel based metrics both. This paper is concentrated on effectively utilization of measure visual similarity for local feature based representation. The important role between feature pairs and their reconstruction was played by kernel. Normally, the Euclidean distance or its derivatives are used to define the kernel. The novel efficient local coding based matching kernel (LCMK) method is presented using the advancement in feature coding techniques. This technique has advantages of both kernel based metrics and BoV. This method is scalable visual matching for large scale image sets operations. It also supports liner computational complexity. This paper used famous dataset like Caltech101, PASCAL VOC 2007, Caltech256 and PASCAL VOC 2011 datasets.

7) Boosted Cross-Domain Dictionary Learning for Visual Categorization

This paper [27] utilizes labeled data from other visual domains as the auxiliary knowledge for enhancing the original categorization learning system is presented. This paper extends the idea of classical AdaBoost. It transfers AdaBoost by integrating an auxiliary domain data representation updating mechanism into the original iterative weight updating mechanism of both algorithms. This framework works better with a learned domain-adaptive dictionary pair. Because of that both the auxiliary domain data representations and their distribution are optimized to match the target domain. While simultaneously generating some artificial data in the auxiliary domain with iteration. The system gives more credits to "similar" auxiliary domain samples. This paper showed the effectiveness of the current method on multiple transfer learning scenarios. The used image datasets were Caltech-101 and UCF YouTube.

Comparative analysis of Accuracy

Table 2: Comparison results on Scene -15 dataset

Method	Average
BOW(4000)	82.53±0.34
84706	82.33±0.64
SPM	81.4±0.49
BOW(400)	81.31±0.43
LLC	81.09±0.43
BOW(1000)	80.91±0.38
soft-coding	76.67±0.39

The comparative analysis of accuracy is performed in two part. Firstly, The accuracy comparision is done in Table 1 to

Table 5 for different image datasetlike Scene-15, Caltech 101, Caltech 256, etc.

The Table 6 demostrates overall accuracy comprasion for the various methods.

Table 3: Comparision results on Caltech 256 dataset

Method	**Average Precision (30 iterations)**	**Average Precision (60 iterations)**
ScSPM [25]	34.02±0.35	40.14±0.91
KSPM [6]	29.51±0.52	-
KC [26]	27.17±0.46	-
LSPM	15.45±0.37	16.57±1.01
ResFeats-152 + PCA-SVM [28]	79.5	82.1
ResFeats-152 + sCNN [28]	78	81.9
ResFeats-50 + Scnn [28]	75.4	79.3
Zeiler & Fergus [29]	70.6	74.2
Bo et al. [28]	48	55.2
Sohn et al. [30]	42.1	47.9
kspm	34.1	-
ResFeats Res5c [25]	-	79.3
Chatfield et al. [31]	-	77.6

Table 4: Comparison results on MIT-indoor67

Method	**Average Precision**
Cimpoi et al. [32]	81
ResFeats-152 + PCA-SVM [25]	75.6
Hayat et al.[33]	74.4
ResFeats-152 + sCNN classifier [25]	73.7
liu et al. [34]	71.5
Azizpour et al. [35]	71.3
ResFeats Res5c [25]	71.1
ResFeats-50 + sCNN classifier [25]	71.1
khan et al.[36]	70.9
zhou et al. [37]	70.8
Gong et al. [38]	68.9
Razavian et al.	58.4
Deep Convolutional Neural Networks [11]	68.3
l2 Normlization + Selective Search + Spatial Pyramid [24]	68.29
Places-CNN feature	68.24
Selective Search + Spatial Pyramid	68.04
Entire Image CNN Features	59.95
ImageNet-CNN feature [12]	56.79
DPM+GIST-color+SP	43.1
Object Bank	37.6

Table 5: Comparison results on MLC dataset

Method	**Average**
ResFeats-152 + PCA-SVM	80.8
ResFeats-152 + sCNN classifier	80
ResFeats-50 + sCNN classifier	78.8
Mahmood et al.	77.9
ResFeats Res5c [25]	76.8
Khan et al.	75.2
Beijbom et al.	74

Table 6: Comparision results on Caltech 101 dataset for 30 iterations

Method	**Average Precision**	**Method**	**Average Precision**
LCMK	80.2	LLC	71.25±0.98
O2P	79.3	BOW(4000)	71.24±0.26
FisherVector	77.8	Zhang et al.	66.2±0.5
BoV	76.9	KSPM	64.4±0.8
HMP	76.8	SPM	64.4±0.8
LLC	73.4	KC [26]	64.14±0.88
Sparse Coding	73.2	soft-coding	64.1±1.2
NBNN [18]	73	KSPM [6]	63.99±0.88
MKL-GL (1)	85.4±0.4	He et al.	93.4
VSKL	85.3±0.6	ResFeats-152 + sCNN	92.6
GMKL [39]	84.8±0.7	ResFeats Res5c	91.8
SimpleMKL [39]	84.6±0.5	ResFeats-50 + Scnn	91.8
level-MKL [39]	84.4±0.4	Chatfield et al.	88.4
MKL-SIP (1) [39]	83.9±0.7	Zeiler & Fergus	86.5
MKL-GL (2) [37]	80±0.6	Bo et al.	81.4
MKL-GL (4) [39]	80±0.6	ResFeats Res5c	75.4
MKL-SMO (2) [39]	79.3±0.9	ScSPM [25]	73.2±0.54
MKL-SIP (2) [39]	79.1±0.6	ML+CORR	69.6
MKL-SMO (4) [39]	79.0±0.5	84706	72.78±0.32
MKL-SIP (4) [39]	77.5±0.5	BOW(400)	72.02±0.26

As show in Table. 2, BOW(4000) provide the supierer accuracy for Scene 15 dataset. The Table 3. is clearly shows that ResFeats-152 + PCA-SVM, ResFeats-152 + sCNN and ResFeats-50 + sCNN provide the better accuracy in neutral network techquies for the Caltech 256 dataset. From Table. 4, The Cimpoi et al. provide the best accuracy for MIT-indoor67 dataset. The ResFeats-152 + PCA-SVM provide the best accuracy for MLC dataset which shows by Table. 5.

As show in Table. 6, the method namely, GMKL, He et al. and KSPM was give the best accuracy for the Caltech 101

dataset. While in neutral network techquies, ResFeats-152 + PCA-SVM, ResFeats-152 + sCNN and ResFeats-50 + sCNN provide the better accuracy.

Table 7: Methods which give overall good accuracy

Method	Accuracy
ResFeats+ PCA-SVM	94.7
ResFeats-50 + sCNN	93
ScSPM	91.8
Sparse Coding	87.45
GMKL	85.4
SPM	81.4
BOW	80.91
Adaptive Aggregating Multiresolution	80.2
LCMK	80

From Table 7, the CNN based ResFeats technique provide highest accuracy then other techquies.

DISCUSSION

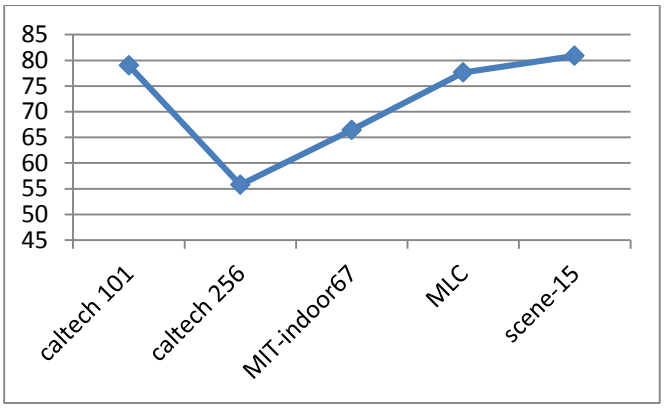


Figure 7: Average accuracy chart for various dataset

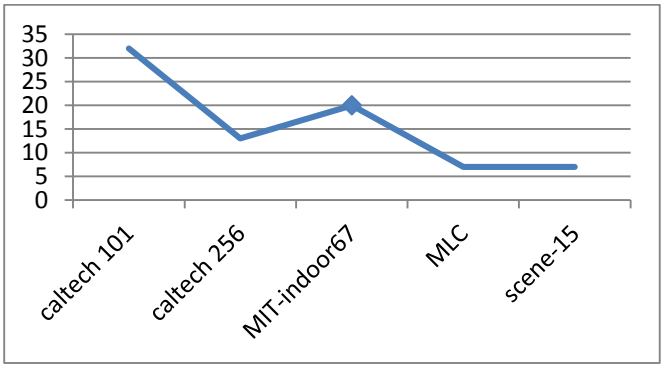


Figure 8: Number of mehods for various dataset

The Fig. 7 show the average accuracy which were achived for various dataset. The Fig. 8 show the total number of various methods whichare analysed for different dataset.

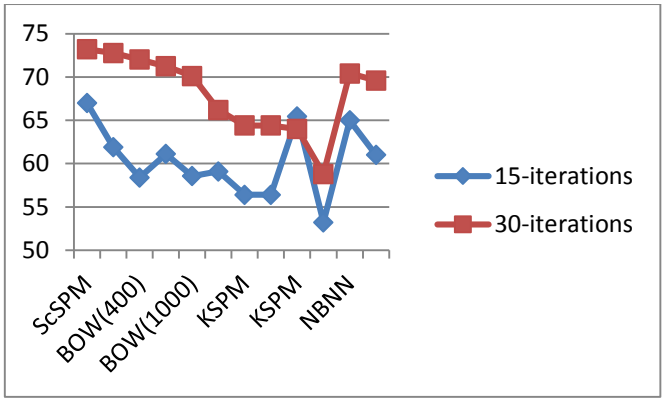


Figure 9: Accuracy chart for Caltech 101dataset

As shown in above Fig. 10, it is clearly shows that method with 60 iteration were give better output then the method with 30 iteration. The ResFeats-152 + PCA-SVM, ResFeats-152 + sCNN and ResFeats-50 + sCNN method were provide good accuracy then rest of the methods.

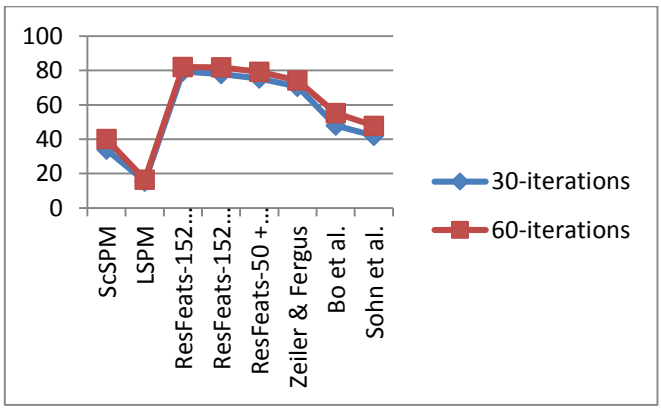


Figure 10: Accuracy chart for Caltech 256 dataset

Figure 11: Average Accuracy chart for various methods

It is clearly shows from Fig. 11 that "ResFeats+ PCA-SVM" method provide the superior classification accuracy then rest of the method.

The MLP has adaptive learning on given training data and not have any assumation for underlying probability denstiy functions which make it best suitatable candidate for gesture

recognition. It cannot be programmed to perform a specific task. The Deep neural network is known to cause overfitting and saturation in accuracy which problem was overcome by ResFeats. The SVM with simple geometric interpretation and give a sparse solution has more advantage then ANN. The first weaknesses of ANNs is that converge on local minima rather than global minima. The second weakness is often overfit if training goes on too long. The SVM overcome these weakness of ANN. The main advantage of ANN is that it has many outputs. So, for the n-ary classifier, ANN is more appropriate then SVM. ANNs use empirical risk minimization, whilst SVMs use structural risk minimization.

CONCLUSION

This paper deals with the scene classification which is the important part of computer vision application start from surveillance application to more complex self driving car. The ScSPM, SVM, deep neural network and various hybrid techniques are compared with average accuracy. The various dataset used for the same are Caltech-101, Caltech-256, MLC, SUN, MIT-indoor67, Scene-15. The highest number of techiniues applied dataset is Caltech-101. The CNN based ResFeat technique provide the superior result then rest of its peer technique.

ACKNOWLEDGMENTS

Authors would like to thank management of The Charotar University of Science and Technology (CHARUSAT), Changa for providing technological support to carry out research at the institute.

REFERENCE

[1] McCulloch, Warren S., and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity." *The bulletin of mathematical biophysics* 5.4, pp.115-133, 1943.

[2] Kuppuswamy, S. and Panchanathan, B., Similar Object Detection and Tracking in H. 264 Compressed Video Using Modified Local Self Similarity Descriptor and Particle Filtering.

[3] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014.

[4] Bhatt, M. and Patalia, T., Neural Network Based Indian Folk Dance Song Classification Using MFCC and LPC.

[5] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference,* IEEE, 2009.

[6] Li, Li-Jia, et al. "Object bank: A high-level image representation for scene classification & semantic feature sparsification." *Advances in neural information processing systems,* 2010.

[7] Herouane, O., Moumoun, L., Gadi, T. and Chahhou, M., A Hybrid Boosted-SVM Classifier for Recognizing Parts of 3D Objects.

[8] Available on: http://ces.iisc.ernet.in/hpg/envis/Remote/section27.htm

[9] Available on: https://www.semanticscholar.org/paper/SUN-database%3A-Large-scale-scene-recognition-from-to-Xiao-Hays/313c782f18bb01933668dce56003553b49d1fc44

[10] Quattoni, Ariadna, and Antonio Torralba. "Recognizing indoor scenes." *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009.

[11] Lazebnik, Svetlana, Cordelia Schmid, and Jean Ponce. "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories." *Computer vision and pattern recognition, 2006 IEEE computer society conference on*. Vol. 2. IEEE, 2006.

[12] Wang, Y. and Wu, Y., Scene Classification with Deep Convolutional Neural Networks.

[13] Wang, Panqu, and Garrison W. Cottrell. "Basic Level Categorization Facilitates Visual Object Recognition." *arXiv preprint arXiv:1511.04103*, 2015.

[14] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.

[15] Fei-Fei, Li, and Pietro Perona. "A bayesian hierarchical model for learning natural scene categories." *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. Vol. 2. IEEE, 2005.

[16] Quelhas, Pedro, et al. "Modeling scenes with local descriptors and latent aspects." *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*. Vol. 1. IEEE, 2005.

[17] Bosch, Anna, Andrew Zisserman, and Xavier Muñoz. "Scene classification using a hybrid generative/discriminative approach." *IEEE transactions on pattern analysis and machine intelligence* 30.4, pp.712-727 , 2008.

[18] Boiman, Oren, Eli Shechtman, and Michal Irani. "In defense of nearest-neighbor based image classification." *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008.

[19] Jurie, Frederic, and Bill Triggs. "Creating efficient codebooks for visual recognition." *Computer Vision,*

2005. ICCV 2005. Tenth IEEE International Conference on. Vol. 1. IEEE, 2005.

[20] Elad, Michael, and Michal Aharon. "Image denoising via sparse and redundant representations over learned dictionaries." *IEEE Transactions on Image processing* 15.12, pp.3736-3745, 2006.

[21] Moosmann, Frank, William Triggs, and Frederic Jurie. "Randomized clustering forests for building fast and discriminative visual vocabularies", 2006.

[22] Yang, Liu, et al. "Unifying discriminative visual codebook generation with classifier training for object category recognition." *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008.

[23] Uijlings, Jasper RR, et al. "Selective search for object recognition." *International journal of computer vision* 104.2 , pp.154-171, 2013.

[24] Yang, Jianchao, et al. "Linear spatial pyramid matching using sparse coding for image classification." *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009.

[25] Van Gemert, Jan C., et al. "Kernel codebooks for scene categorization." *European conference on computer vision*. Springer, Berlin, Heidelberg, 2008.

[26] Song, Y., McLoughlin, I.V. and Dai, L.R., 2014. Local coding based matching kernel method for image classification. *PloS one*, *9*(8), p.e103575.

[27] Zhu, Fan, Ling Shao, and Yi Fang. "Boosted cross-domain dictionary learning for visual categorization." *IEEE Intelligent Systems* 31.3, pp.6-18, 2016.

[28] Bo, Liefeng, Xiaofeng Ren, and Dieter Fox. "Multipath sparse coding using hierarchical matching pursuit." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013.

[29] Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." *European conference on computer vision*. Springer, Cham, 2014.

[30] Sohn, Kihyuk, et al. "Efficient learning of sparse, distributed, convolutional feature representations for object recognition." *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011.

[31] Chatfield, Ken, et al. "Return of the devil in the details: Delving deep into convolutional nets." *arXiv preprint arXiv, 1405.3531*, 2014.

[32] Cimpoi, Mircea, et al. "Deep filter banks for texture recognition, description, and segmentation." *International Journal of Computer Vision* 118.1, pp. 65-94, 2016.

[33] Hayat, Munawar, et al. "A spatial layout and scale invariant feature representation for indoor scene classification." *IEEE Transactions on Image Processing* 25.10, pp. 4829-4841, 2016.

[34] Liu, Lingqiao, Chunhua Shen, and Anton van den Hengel. "The treasure beneath convolutional layers: Cross-convolutional-layer pooling for image classification." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[35] Azizpour, Hossein, et al. "From generic to specific deep representations for visual recognition." *CVPRW DeepVision Workshop, June 11, 2015, Boston, MA, USA*. IEEE conference proceedings, 2015.

[36] Khan, Salman H., et al. "Cost-Sensitive learning of deep feature representations from imbalanced data." *IEEE transactions on neural networks and learning systems*, 2017.

[37] Zhou, Bolei, et al. "Learning deep features for scene recognition using places database." *Advances in neural information processing system,* 2014.

[38] Gong, Yunchao, et al. "Multi-scale orderless pooling of deep convolutional activation features." *European conference on computer vision*. Springer, Cham, 2014.

[39] Bucak, Serhat S., Rong Jin, and Anil K. Jain. "Multiple kernel learning for visual object recognition: A review." *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 36.7, pp. 1354-1369, 2016.