IJCNN 2019. International Joint Conference on Neural Networks. Budapest, Hungary. 14-19 July 2019

Scene Recognition via Object-to-Scene Class Conversion: End-to-End Training

Hongje Seong, Junhyuk Hyun, Hyunbae Chang, Suhyeon Lee, Suhan Woo and Euntai Kim*
School of Electrical and Electronic Engineering
Yonsei University
Seoul, South Korea
{hjseong, jhhyun, chang0977, hyeon93, wsh112, etkim}@yonsei.ac.kr

***Abstract*—When a person recognize the scene of an image, contextual understanding from its environmental elements is necessary. These environmental elements are variant and require comprehensive understanding of various situations. Especially, objects are frequently used as environmental elements related with scene. In this paper, we suggest a score level Class Conversion Matrix (CCM) for scene recognition with a great focus on relationship between objects and scene. A lot of existing methods have already build scene recognition systems with consideration of close relationship between object and scenes. However, most of these methods are using the object features directly without any conversions or reconstructions, and it lack confirmation whether these object features are helpful to recognize scenes correctly. To solve this problem, CCM, a matrix converting object feature to scene feature, is suggested. Moreover, CCM can be implemented with neural network layer and end-to-end trainable. Extensive experiments on Places 2 dataset demonstrate the effectiveness of our approach, when it is applied to the existing deep convolutional neural network architectures. The code is available at https://github.com/Hongje/Class_Conversion_Matrix-Places365**

***Index Terms*—scene recognition, class conversion matrix, end-to-end trainable**

I. Introduction

Scene recognition has been studied in computer vision and robotics field. In many computer vision studies, it is considered as a categorization problem [1], [2]. However, these methods do not contain any analysis of scene characteristics. In robotics area, scene recognition techniques can help robot localization [3].

When people recognize a scene image, they are affected by scene characteristics as well as environmental elements. This paper focuses on the objects as environmental elements of a scene. For example, objects like blackboard, desk, chair are expected as environmental elements in a classroom scene. Furthermore, if a person standing in front of blackboard is considered as a teacher while sitting on a chair as a student in the same scene. With this kind of idea, we derived a conclusion; relationship between scene and object can improve scene recognition performance with CCM.

The convolutional neural network (CNN) structure is commonly used for classification problem since it won ImageNet large scale visual recognition challenge (ILSVRC) in 2012 [4], [5]. Because so many parameters have to be trained, deep CNN structure requires a huge amount of training data. One of the success factors of the CNN structure at ILSVRC 2012 is ImageNet dataset that provides enough amount of training data. In scene recognition, there is a Places 2 [6] dataset, which also provides as large amount of training data as the ImageNet dataset, which is sufficient to train the CNN structure. Therefore, we applied CNN structure to object and scene categorization problem and extracted object and scene features using each CNNs trained on ImageNet and Places 2 dataset respectively.

This paper proposes a new fusion method of high level object and scene features. The existing scene recognition systems [7]–[12] focused on how to extract the useful object features only, but there were not any consideration about how to integrate these features. We extract object and scene scores from ImageNet and Places CNN respectively and combine those scores with CCM that represents the relation between object and scene. Since CCM uses matrix multiplication, there is no concern on back-propagation [13]. So, CCM is not only a neural network layer but also a end-to-end trainable module.

The whole system of scene recognition using the CCM is shown at Fig. 1. This system requires additional two modules able to get both of object and scene scores. In this paper, two simple and powerful CNNs are used to extract object and scene scores respectively. However CCM is not limited to CNN, and another modules [14]–[17] are also available because CCM only need scores.

The remaining sections are organized as follows: Section II describes the briefly reviews for related studies, and Section III shows the proposed approach. Section IV illustrates the experiments with the Places 2 dataset, followed by the conclusion in Section V.

II. Related Work

In this section, the briefly reviews of scene recognition related with our research is described. The topics are divided into two parts: (II-A) scene recognition, and (II-B) objects in scene recognition..

A. Scene Recognition

There have been many researches for scene recognition using image. The methods using hand crafted features [18],

*Corresponding author

978-1-7281-2009-6/$31.00 ©2019 IEEE

Personal use is permitted, but republication/distribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

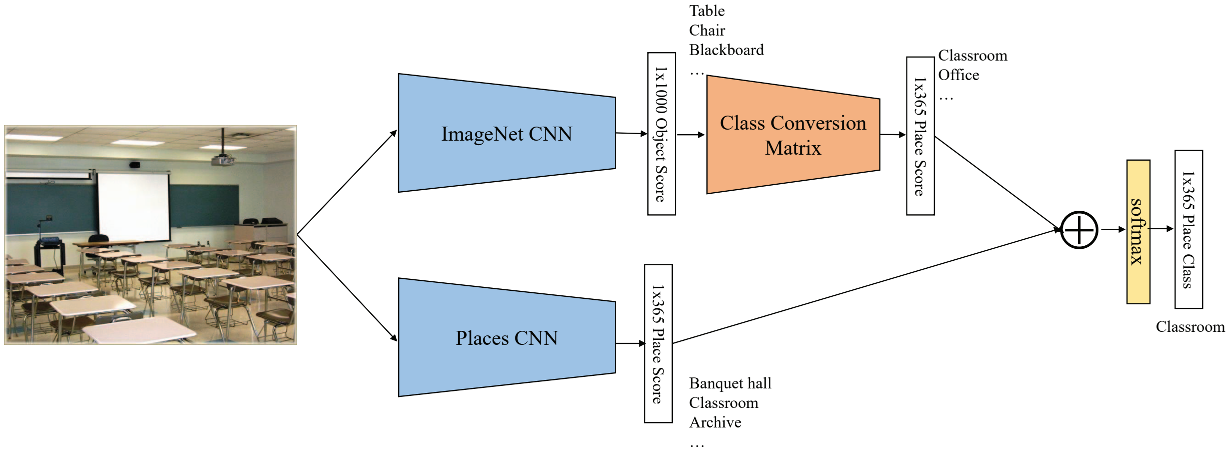


Fig. 1. An example of the proposed scene recognition system using score level CCM with object score on Places 2 dataset. The symbol ⊕ is element-wise sum. The CCM not violate the rule of back-propagation. Therefore this is end-to-end trainable system. In addition, this system propose a fusion method of score that is high level feature. Thus, it can be adopted any scene recognition system that extract object and scene score or other environmental elements that contain scene information.

[19] used filters to create features representing the scene well. The methods using a CNN are appeared after AlexNet [5]. In general, the models with outstanding performance on object classification have better performance in scene recognition. The method using support vector machine (SVM) [20] with CNN [7]–[12], [21]–[28] are frequently used. These methods adopt the CNN structure for extracting features of an image and classify scene categories with SVM. In this case, it is hard to train in end-to-end because CNN and SVM are employed at the same time. Thus, it is better to train the CNN with large scene dataset first and extract the features without classification layer from the CNN. And then, the features from pretrained CNN are used in SVM training.

A single CNN methods [1], [2] do not focus on scene recognition but use it to show the CNN architecture work well in general. Even though these methods can be trained in end-to-end, they do not consider about scene characteristics. It means that there are a lot of opportunity to improve the scene recognition performance in end-to-end structures.

B. Objects in Scene Recognition

There are many scene recognition approaches using object information for scene recognition [7], [9], [10], [12]. Most of these methods train the scene and object modules with Places 2 [6] and ImageNet dataset [4]. In commonly, extracting object and scene features using each module, and combine these features using sum or concatenate method. Then, combined feature is used at SVM training. Even though these kinds of systems can be run as end-to-end, training in end-to-end system is not possible.

Most of these methods are focusing on how to extract meaningful object features. The methods simply concatenating or adding features display that using object information enhanced the scene recognition performance [7], [9], [10], [12]. All of these approaches are concentrating on ways of feature extraction without any consideration on how to combine object and scene features.

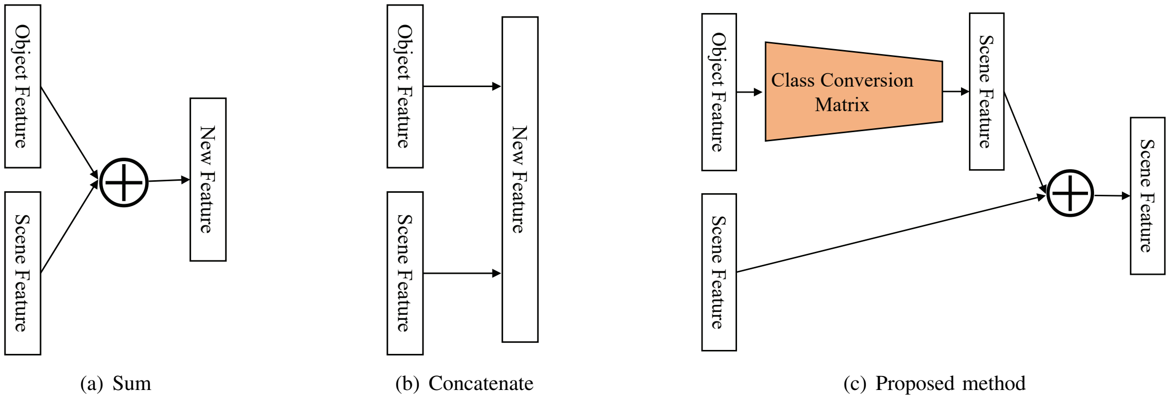
III. Proposed Approach

In this section, our proposed method is described. For scene recognition, we focused on the analysis of correlation of object and scene.

A. Motivation

Even though there was a great improvement of various CNN based methods of scene recognition, most of existing scene recognition system did not focus on how to fuse the object and scene features. This makes hard to construct end-to-end trainable system. Fig. 2 shows the differences between previous feature fusion methods and ours. Simple adding with the same dimension after feature extractions is Fig. 2(a). It does not preserve the original information but creates a new feature. Therefore, many existing methods [7], [9]–[12] uses concatenation as shown in Fig. 2(b). Even if this method preserves the original information and earns better performance, it is hard to expect the meaningful usage of object feature, too. In addition, it is able to expect that the performance improvement by more amount of features rather than object information.

Different with the two methods mentioned already, we wanted such a system that is able to use object features more usefully. Furthermore, designing an easier architecture with end-to-end train is desired. As a result, we suggest a module can convert object features to more meaningful one. As displayed in Fig. 2(c), converted feature is added to original scene feature. Even though we added the features to fuse the object and scene features together, it has different meaning from Fig. 2(a)'s. Fig. 2(a)'s method is meaningless due to adding two totally different features. At the same time, the method used in Fig. 2(c) is adding scene features after



(a) Sum (b) Concatenate (c) Proposed method

Fig. 2. The feature level fusion methods of object feature and scene feature. (a) is a method to adding the features. It can be applied only when the object and scene features have same dimension. (b) is a method of increasing the amount of information by concatenating object and scene features. (c) is a proposed method that convert a object feature to a scene feature and combine features with similar characteristic.

converting the information that scene features cannot describe to object features.

This approach has additional computations due to increase parameters, but is not a big cost because it is possible to end-to-end training and applied to high level features.

B. Class Conversion Matrix

The scene recognition system using CCM is shown at Fig 1. High level features like score are the input of CCM. Object score is extracted from ImageNet CNN. However, it can be substituted to better ways to extract object information.

$\mathbf{X}$ is an input score vector while $\mathbf{W}$ and $\mathbf{b}$ are the trainable parameters of CCM. Then if the output score vector is $\mathbf{Y}$, CCM is computed as follow:

$$\mathbf{Y} = \mathbf{W}\mathbf{X} + \mathbf{b} \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^n$ is an object score of high level feature that helps scene recognition. $\mathbf{W} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$ are the trainable parameters and $\mathbf{Y} \in \mathbb{R}^m$ is a scene score related to $\mathbf{X}$. n is dimension of an input vector and m is output vector's. If ImageNet dataset [4] is used to train object module, $n = 1000$. And, if Places 2 dataset [6] is used to train scene module, $m = 365$. CCM can be applied to not only the different datasets but also the same datasets. For example, the parameters of CCM $\mathbf{W} \in \mathbb{R}^{m \times m}$ and $\mathbf{b} \in \mathbb{R}^m$ with the Places 2 dataset format, output $\mathbf{Y} \in \mathbb{R}^m$ and input $\mathbf{X} \in \mathbb{R}^m$ with the Places 2 dataset formats are also configurable.

After CCM computation, activation follows e.g. ReLU, linear, etc.

CCM can use batch normalization (BN) [29] with matrix multiplication. BN used CCM is computed as

$$\mathbf{Y} = \mathrm{BN}\left(\mathbf{W}\mathbf{X}\right). \tag{2}$$

Bias $\mathbf{b}$ in the (1) is replaced by batch normalization. Due to the characteristics of BN, scene score $\mathbf{Y}$, transformed from object score $\mathbf{X}$, is normalized with using mean and variance of total score. Therefore, scene score $\mathbf{Y}$ is not biased to specific scene class. It can reduce the effect of useless object information on scene score, which means, we can expect better scene recognition performance. CCM with BN (CCM-BN) also can be replaced by neural network layer, an end-to-end trainable structure.

The two types of CCM (1), (2) are possible to calculate error of back-propagation, so it can express to neural network layer as Fig. 1. Therefore, the whole system that contains conversion matrix will be end-to-end trainable.

C. Fusion Level

In this paper, we present fusion methods at two levels. One is fusion at score level and the other one is fusion at feature level.

Score level fusion is a method of fusing class score vectors. Fig. 1 represent the score level fusion method using CCM. At object score vector, 1×1000 vector is used because ImageNet has 1000 classes. Similarly, 1×365 score vector is used at scene score vector because Places 2 has 365 classes. It is not necessary to use a neural network structure to extract score vectors. CCM can also be applied to score extracted from SVM, but in this case, end-to-end training is not available.

Feature level fusion is a method of fusing the input vector of last fully connected layer. In this case, each CNN architecture has a fixed dimension of feature vector. Therefore, many fusion methods use this method because it is easy to extract object and scene information in the same dimension. The figures for these schemes are shown in Fig. 2.

IV. Experiments

In this section, our experiments and results are described. We evaluated our methods with Places 2 dataset [7], a large scale scene recognition dataset.

A. Dataset

We designed experiments based on Places 2 dataset [6]. This dataset is a new version of the Places dataset [30], which consists of about 2.4 million images and 205 scene categories. Places 2 dataset provides two types of datasets: Places365 challenge dataset contains 8 million images with 365 categories and the other one, Places365-Standard dataset,

TABLE I
ACCURACY(%) OF CLASS CONVERSION MATRIX(CCM) WITH SEVERAL CNN ARCHITECTURE ON PLACES 2 DATASET

Fusion Level	Method	AlexNet		ResNet-18		ResNet-50		DenseNet-161	
		top-1 acc	top-5 acc	top-1 acc	top-5 acc	top-1 acc	top-5 acc	top-1 acc	top-5 acc
	Original (Provided by Places 2)	47.551	77.984	53.693	83.778	54.767	84.932	56.132	86.121
	Re-implementation	48.452	78.975	54.721	84.836	56.000	86.016	56.605	86.438
Score Level	CCM	49.027	**79.501**	54.847	**84.981**	56.093	**86.307**	56.471	86.592
	CCM-BN	**49.049**	79.452	**54.910**	84.975	**56.195**	86.293	**56.682**	**86.671**
Feature Level	Sum	48.337	78.986	54.216	84.274	56.266	86.274	55.729	85.866
	Concatenate	48.784	79.406	55.052	85.126	56.408	86.405	56.723	86.663
	CCM	48.956	79.685	54.918	85.074	56.452	86.447	56.616	86.751
	CCM-BN	48.995	79.537	54.986	85.129	56.395	86.460	56.655	86.745
	CCM with ReLU	**49.258**	**79.830**	54.978	85.074	56.474	**86.515**	**56.822**	**86.921**
	CCM-BN with ReLU	49.077	79.625	**55.066**	**85.164**	**56.567**	86.499	56.712	86.792

includes 1.8million images with the same number of categories. Compared to the 1.28 million training images provided by the ImageNet 2012 dataset [4], which contributed to the development of CNN, the amount of Places 2 datasets provides a sufficient amount to train CNN. There are other datasets, too: MIT indoor 67 [31], SUN 397 [32], Scene 15 [33]. MIT indoor 67 dataset [31] consists of 15,620 images and 67 scene categories while SUN 397 [32] is composed of 108,754 images with 397 scene categories. And Scene 15 [33] has 4485 gray images of 15 scene categories. Since all those three datasets provide less images than Places 2 datasets does, it is hard to train CNN with those. Many CNN based scene recognition methods using these datasets are used only features extracted by CNN trained on Places 2 dataset without additional training on target datasets.

There is a Places365-pretrained CNN trained with Places365-Standard dataset [34]. It is used as a baseline places CNN in this paper. To achieve a fair experiment with the same environment, the experiments would go through with Places365-Standard dataset.

B. Implementation Details

We extracted object features with ImageNet CNN trained with ImageNet datset [4]. Similarly, scene features are gained from pretrained Places CNN. And build a scene recognition system with object information as displayed at Fig 1. Then, we applied transfer learning method [35] to conserve the object information acquired from ImageNet CNN.

The training hyper parameters adopted in this experiments are: mini-batch size is 256, initial learning rate is 0.01, and decay the learning rate divide by 10 every 30 epochs. However, not the same mini-batch sizes are employed at all models. To increase the training speed, learning rates were varied according to the linear scaling rule [36] whenever mini-batch sizes are changed. Additionally, we used two types of optimizers: stochastic gradient descent (SGD) with momentum of 0.9, and Adam [37] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. After training with both of them separately, we chose our experiment results based on validation performance. We measured top-1 and top-5 classification accuracy with 224×224 center cropped images. For parameter initialization, we used ImageNet CNN provided from Pytorch [38] and Places CNN from officially released Places 2 dataset [34]. We confirmed the performance elevation with Places 2 fine-tuning. Thus, re-implementation model is adopted and its performance can be checked at Table I.

C. Experiment Results

Table I shows top-1 and top-5 validation accuracy with CCM. We applied AlexNet [5], Resnet-18, ResNet-50 [39], and DenseNet-161 [40] to show CCM can work well generally, not restricted to some specific models. Places CNNs for AlexNet, Resnet and DenseNet can be gained from [34]. We tested the validation accuracy without any additional training and it is displayed at the first row (named Original) of Table I.

Table I also shows the performance of CCM fused at score level. When CCM is applied, the model outperformed the baseline. The object information filtered with CCM can help to solve the scene ambiguity not solvable with single scene dataset. IV-D shows how our proposed method help the scene recognition with object information.

Moreover, there are some experiments comparing the proposed methods to sum and concatenation methods. To apply sum and concatenation methods, new classifier trained on new fused features is necessary as shown in Fig. 2(a) and (b). Accordingly, we proceeded an experiment fusing features extracted before last fully connected layer. After that, last fully connected layer is trained. This experiment is displayed in Table I's feature level. CCM applied method outperformed other existing methods. In the case of summation method, deep networks like ResNet-50 and DenseNet-161 earned better performance than baseline, but shallow networks of AlexNet and ResNet-18 gave worse results. Concatenation gave the higher performances from all the models but lower than the proposed method. Therefore, our proposed method uses the object information to improve scene recognition performance which is different from simple summation and concatenation methods.

D. Analysis of Relationship Between Scene and Object

We can analyze the relationship between scene and object with weights of CCM. If given object appears frequently in provided scenes, CCM would have higher weights but lower ones in the opposite case.

TABLE II
TOP-5 WEIGHTS OF CLASS CONVERSION MATRIX

Weight Value	Places 2	ImageNet
0.09625	field/wild	ox
0.09385	desert/sand	Arabian camel
0.09278	field/wild	yellow lady's slipper
0.09219	gas station	gas pump
0.09043	desert road	car mirror

TABLE III
BOTTOM-5 WEIGHTS OF CLASS CONVERSION MATRIX

Weight Value	Places 2	ImageNet
-0.09954	field/wild	broom
-0.09859	field/wild	can opener
-0.09716	desert/vegetation	snowplow
-0.09403	field/wild	tricycle
-0.09350	desert road	coral reef

Table II and III show the analysis of top-5 and bottom-5 CCM weights respectively to convince that object score after CCM is effective to help scene recognition. Field/wild class is the most influenced class by object information among whole classes and this class is expected to have many objects inside the scenes. The results shown in Table II are reasonable objects that may exist in the given scene classes. Similarly, the combinations of scene and object classes shown in Table III convince that they are objects that should not exist in given scene.

V. CONCLUSION

In this work, we proposed CCM for scene recognition. We experimented with several CNN architectures and showed that our proposed method is valid for scene recognition. Our proposed module is available to be trained as end-to-end system, so there is not any limitation to use our module for designing a neural network. In addition, even if it is not a neural network structure, it can be applied to any system that can extract any scene recognizing information.

The presented method is focus only on a fusion method of features between scene and other information without considering the generation of information. If the information is very helpful in recognizing the scene, it can expect enormous performance improvement with our proposed fusion method. Therefore, we will improve the performance by focus on research in a method of extracting a helpful feature in scene recognition later.

ACKNOWLEDGMENT

This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT (NRF-2017M3C4A7069370).

REFERENCES

[1] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[2] L. Shen, Z. Lin, and Q. Huang, "Relay backpropagation for effective learning of deep convolutional neural networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[3] S. Garg, A. Jacobson, S. Kumar, and M. Milford, "Improving condition- and environment-invariant place recognition with semantic place categorization," in *Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 6863–6870.

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

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1097–1105.

[6] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2018.

[7] N. Sun, W. Li, J. Liu, G. Han, and C. Wu, "Fusing object semantics and deep appearance features for scene recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[8] L. Wang, S. Guo, W. Huang, Y. Xiong, and Y. Qiao, "Knowledge guided disambiguation for large-scale scene classification with multi-resolution cnns," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 2055–2068, 2017.

[9] Z. Zhao and M. Larson, "From volcano to toyshop: Adaptive discriminative region discovery for scene recognition," in *Proceedings of the 26th ACM international conference on Multimedia (ACM MM)*, 2018, pp. 1760–1768.

[10] L. Herranz, S. Jiang, and X. Li, "Scene recognition with cnns: objects, scales and dataset bias," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 571–579.

[11] G.-S. Xie, X.-Y. Zhang, S. Yan, and C.-L. Liu, "Hybrid cnn and dictionary-based models for scene recognition and domain adaptation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 6, pp. 1263–1274, 2017.

[12] X. Cheng, J. Lu, J. Feng, B. Yuan, and J. Zhou, "Scene recognition with objectness," *Pattern Recognition*, vol. 74, pp. 474–487, 2018.

[13] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep., 1985.

[14] P. Gehler and S. Nowozin, "On feature combination for multiclass object classification," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2009, pp. 221–228.

[15] M. Marszalek and C. Schmid, "Semantic hierarchies for visual object recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2007.

[16] S. S. Bucak, R. Jin, and A. K. Jain, "Multiple kernel learning for visual object recognition: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1354–1369, 2014.

[17] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: Binary robust independent elementary features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2010, pp. 778–792.

[18] S. N. Parizi, J. G. Oberlin, and P. F. Felzenszwalb, "Reconfigurable models for scene recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 2775–2782.

[19] M. Pandey and S. Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2011.

[20] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, 1992, pp. 144–152.

[21] A. Bayat and M. Pomplun, "Deriving high-level scene descriptions from deep scene cnn features," in *Image Processing Theory, Tools and Applications (IPTA), 2017 Seventh International Conference on*. IEEE, 2017.

[22] S. Yang and D. Ramanan, "Multi-scale recognition with DAG-CNNs," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1215–1223.

[23] J. de Jesús Rubio, E. Lughofer, J. A. Meda-Campaña, L. A. Páramo, J. F. Novoa, and J. Pacheco, "Neural network updating via argument

kalman filter for modeling of takagi-sugeno fuzzy models," *Journal of Intelligent & Fuzzy Systems*, vol. 35, no. 2, pp. 2585–2596, 2018.

[24] X.-L. Meng, F.-G. Shi, and J.-C. Yao, "An inequality approach for evaluating decision making units with a fuzzy output," *Journal of Intelligent & Fuzzy Systems*, vol. 34, no. 1, pp. 459–465, 2018.

[25] J. de Jesús Rubio, "Stable kalman filter and neural network for the chaotic systems identification," *Journal of the Franklin Institute*, vol. 354, no. 16, pp. 7444–7462, 2017.

[26] M.-Y. Cheng, D. Prayogo, and Y.-W. Wu, "Prediction of permanent deformation in asphalt pavements using a novel symbiotic organisms search–least squares support vector regression," *Neural Computing and Applications*, 2018.

[27] J. de Jesús Rubio, "Sofmls: online self-organizing fuzzy modified least-squares network," *IEEE Transactions on Fuzzy Systems*, vol. 17, no. 6, pp. 1296–1309, 2009.

[28] X.-M. Zhang and Q.-L. Han, "State estimation for static neural networks with time-varying delays based on an improved reciprocally convex inequality," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 4, pp. 1376–1381, 2018.

[29] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2015.

[30] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 487–495.

[31] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 413–420.

[32] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 3485–3492.

[33] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2006, pp. 2169–2178.

[34] B. Zhou, "The places365-cnns for scene classification," https://github.com/CSAILVision/places365, 2017.

[35] S. J. Pan, Q. Yang *et al.*, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[36] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch sgd: training imagenet in 1 hour," *arXiv preprint arXiv:1706.02677*, 2017.

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[38] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS 2017 Autodiff Workshop: The Future of Gradient-based Machine Learning Software and Techniques*, 2017.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[40] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks." in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.