Evaluation of Basic Convolutional Neural Network, AlexNet and Bag of Features for Indoor Object Recognition

Srie Azrina Zulkeflie, Fatin Amira Fammy, Zaidah Ibrahim, and Nurbaity Sabri

Abstract—**This paper evaluates two deep learning techniques that are basic Convolutional Neural Network (CNN) and AlexNet along with a classical local descriptor that is Bag of Features (BoF) with Speeded-Up Robust Feature (SURF) and Support Vector Machine (SVM) classifier for indoor object recognition. A publicly available dataset, MCIndoor20000, has been used in this experiment that consists of doors, signage, and stairs images of Marshfield Clinic. Experimental results indicate that AlexNet achieves the highest accuracy followed by basic CNN and BoF. Furthermore, the results also show that BoF, a machine learning technique, can also produce a high accuracy performance as basic CNN, a deep learning technique, for image recognition.**

Index Terms—**AlexNet, Bag of Features (BoF), Convolutional Neural Network (CNN), indoor object recognition.**

I. Introduction

Image processing and machine intelligence has been implemented and utilize in every aspect of human daily life. Two approaches for computer vision are Machine Learning (ML) and Deep Learning (DL). Detection of significant pattern in data known as automated detection is one of machine learning algorithm [1] while Deep learning (DL) is a great machine learning methodology for overcome a complex problems in image processing, natural language processing, computer vision, and signal processing. One main difference between ML and DL is that the application of ML requires the two phases namely, feature extraction and classification while DL does not separate these two phases. DL has been used for various researches in object recognition such as character recognition [2], herb leaf recognition [3], and face recognition [4]. One of the popular techniques under ML is Bag of Features (BoF) where it has been used in various computer vision applications such as scene character recognition [5], food recognition [6], and vehicle recognition [7].

On the other hand, one of the famous techniques under DL is Convolutional Neural Network (CNN) and AlexNet, a CNN pre-trained model. CNN produces excellent solution which can extract a hierarchical representation of invariant input data transformations and scales [8], which has achieved high accuracy in classifying the grading of palm oil Fresh Fruit Bunch (FFB) ripeness [8]. Besides that, CNN is also capable of producing high accuracy in classifying patients review towards doctors and healthcare services [9]. Meanwhile, AlexNet has proven to obtain excellent performance for ear recognition [10].

BoF with Speeded-Up Robust Features (SURF) and Support Vector Machine (SVM) has been applied to recognize vehicle make and its model [7]. Using a single dictionary, it manages to achieve 95.77% accuracy compared to modular dictionary. This algorithm is able to recognize vehicle under occlusion, non-frontal vision object and object with dim luminescence environment [7]. Scale-Invariant Feature Transform (SIFT), one of the most robust features other than SURF [11], has been implemented on batik image classification [12]. High accuracy results have been achieved using this combination. However, SIFT produce high accuracy only for simple and less noisy background images. Naïve Bayes classifier produces a good result compared to SVM in human detection in video surveillance [13]. This classifier also achieved a high accuracy for human action recognition which is 99.4% [14]. However, it needs an intensive computation operation to perform this classification and the results produced are similar to the result produced using threshold-based system. Besides, this classifier needs a very large number of probability dataset to produce good results [15]. Random forest classifier manages to increase the detection accuracy on wildfire smoke with the implementation of BoF model [16]. However, this classifier requires a long training time due to its complex numeric dataset and known as an unstable algorithm [15].

Indoor object recognition is useful for indoor robot navigation and mobility for visually impaired person [17]. A publicly available dataset called MCIndoor20000 has been constructed for research purposes that consist of doors, signs, and stairs indoor images in a clinic [18].

Since the accuracy performance for object recognition produced by BoF, AlexNet, and basic CNN are outstanding; this paper tends to investigate the accuracy performance of basic CNN, AlexNet, and BoF with SURF alongside SVM for indoor object recognition using MCIndoor20000 dataset. Experiments have also been conducted utilizing CNN and BoF for object classification which effectively increase the classification rate with relatively minimal storage [19]. The rest of the paper is organized as follows: Section II briefly describes the classification methods used for the experiments; Section III explains the dataset and experiments environment;

Manuscript received August 20, 2019; revised October 11, 2019. This work was supported and financially sponsored by the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia.

The authors are with the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia (e-mail: sriesa@gmail.com, fatinamirarikzan@gmail.com, zaidah@tmsk.uitm.edu.my, nurbaity_sabri@melaka.uitm.edu.my).

doi: 10.18178/ijmlc.2019.9.6.876

Section IV presents the results and discussion of the evaluation and followed by the conclusion as the last section.

II. Classification Methods

A. Basic Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNN) is a widely used tool under deep learning. Fig. 1 shows a basic CNN architecture that consists of several layers of various types that are convolutional layers, activation layers, pooling layers, and ends with one or fully connected layers [20].

Each convolution layer consists of number of kernel which produces the same numbers of features maps. It works by sliding the kernels with a particular receptive field over the feature maps from the previous layer. Each feature map that is computed is characterized by several hyper-parameters such as the size and depth of the filters, the stride between filters and the amount of zero-padding around the input feature map [21]. Pooling layers can be applied in order to cope with translational variances and to decrease the size of the feature maps [22]. They proceed by sliding a filter through the feature maps and output the highest or average value. This process depends on the selection of pooling, in each sub-region. The function of Rectified Linear Unit (ReLU) is for a nonlinear or activation layer that is applied to a feature map after each convolutional layer because of the computational efficiency and the alleviation of the vanishing gradient problem [23]. The fully connected layers typically are the last few layers of the architecture. The number of classes to be recognized contains the same number of neurons produce by the final fully connected layers of CNN architecture.

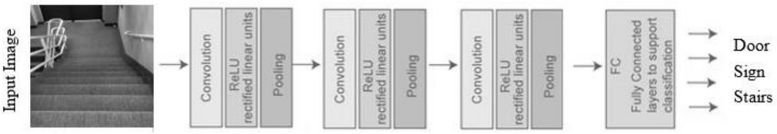
Input Image
Convolution
ReLU rectified linear units
Pooling
Convolution
ReLU rectified linear units
Pooling
Convolution
ReLU rectified linear units
Pooling
FC
Fully Connected layers to support classification
Door
Sign
Stairs

Fig. 1. CNN architecture that consists of various types of layers [5].

TABLE I: Architecture of AlexNet

Layer	Layer Name	Layer Type	Layer Details
1	'data'	Image Input	227×227×3 images with 'zerocenter' normalization
2	'conv1'	Convolution	96 11×11×3 convolutions with stride [4 4] and padding [0 0 0 0]
3	'relu1'	ReLU	ReLU
4	'norm1'	Cross Channel Normalization	cross channel normalization by 5 channels per element
5	'pool1'	Max Pooling	3×3 max pooling with stride [2 2] and padding [0 0 0 0]
6	'conv2'	Convolution	256 5×5×48 convolutions with padding [2 2 2 2] and stride [1 1]
7	'relu2'	ReLU	ReLU
8	'norm2'	Cross Channel Normalization	cross channel normalization with 5 channels per element
9	'pool2'	Max Pooling	3×3 max pooling with stride [2 2] and padding [0 0 0 0]
10	'conv3'	Convolution	384 3×3×256 convolutions with padding [1 1 1 1] and stride [1 1]
11	'relu3'	ReLU	ReLU
12	'conv4'	Convolution	384 3×3×192 convolutions with stride [1 1] and padding [1 1 1 1]
13	'relu4'	ReLU	ReLU
14	'conv5'	Convolution	256 3×3×192 convolutions with stride [1 1] and padding [1 1 1 1]
15	'relu5'	ReLU	ReLU
16	'pool5'	Max Pooling	3×3 max pooling with padding [0 0 0 0] and stride [2 2]
17	'fc6'	Fully Connected	4096 fully connected layer
18	'relu6'	ReLU	ReLU
19	'drop6'	Dropout	50% dropout
20	'fc7'	Fully Connected	4096 fully connected layer
21	'relu7'	ReLU	ReLU
22	'drop7'	Dropout	50% dropout
23	'fc8'	Fully Connected	1000 fully connected layer
24	'prob'	Softmax	softmax
25	'output'	Classification Output	crossentropyex with 'tench' and 999 other classes

B. AlexNet

In 2012, AlexNet won the ImageNet visual object recognition challenge, i.e. the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [24]. The AlexNet architecture contains eight layers, which consists of five convolutional layers and three fully connected layers. The architecture of AlexNet is shown in Table I. The first convolutional layer performs convolution and max pooling where the filters size used are 11-by-11. The max pooling operations are performed with 3-by-3 filters with a stride size of 2. The second layers with 5-by-5 filter layer also perform the same operations. The max pooling operations are performed with 3-by-3 filters with a stride size of 2. The filter size is 3-by-3 in the third, fourth, and fifth convolutional layers. The max pooling operations are performed with

3-by-3 filters with a stride size of 2 at the fifth layer. Each of the sixth and seventh fully connected layers contains 4,096 neurons. The numbers of classes to be classified by ImageNet dataset consist of 1,000 classes. Therefore the final fully connected layer also contains 1,000 neurons [20]. The ReLU activation function is implements to the first seven layers respectively. A dropout ratio of 0.5 is applied to the sixth and seventh layer. The eighth layer output is finally supplied to a softmax function. Dropout is a regularization technique, being used to overcome the overfitting problem that remains a challenge in a deep neural network [25]. Thus, it reduces the training time for each epoch.

C. *Bag of Features (BoF)*

The most popular approach in image category classification is a Bag of Features technique. It usually referred to as Bag of Words (BoW). The idea of BoW model in computer vision is to consider an image containing of different visual words [6]. Descriptor of an image can be acquired by clustering features of local regions that consists rich information in the images, such as color or texture.

In the image analysis context, an image is represented by the histogram of visual words, which are defined as representative image patches of regularly occurring visual patterns [26]. Since images do not actually contain discrete words, a feature detectors and descriptors such as SURF can be used to build a visual vocabulary of SURF features to represent every image category.

SURF is a robust image detector and descriptor and makes use of integral images and sometimes provides with more than 10% improvements compared to other descriptors [27]. Features from the entire images in the image categories are extracted and visual vocabulary is constructed by decreasing the number of features. It is done through quantization of feature space by applying K-means clustering algorithm. The new and reduced representation of image produces a histogram by calculating the visual word appearances in an image. This histogram will be the reference for actual image classification and training the classifier. The encoded training images from every category are supplied into a classifier training process called by the function that depends on the multiclass linear SVM classifier. SVM performs classification by mapping the input vectors non-linearly into a high dimensional feature space. The recognition performed by construct an optimum separation hyperplane in the space [28]. Fig. 2 illustrates the process flow of BoF.

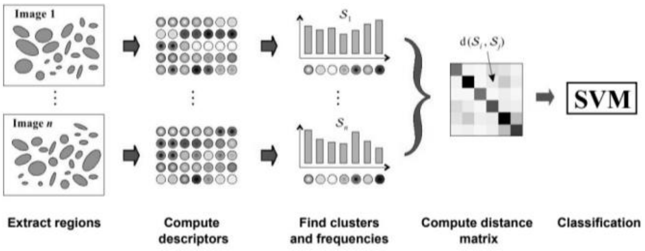

Fig. 2. Bag of features for image classification [29].

III. Experiments

A. *Dataset*

The data that is being used for this experiment is from MCIndoor20000 dataset [18]. Original images acquired from Marshfield Clinic, Marshfield, Wisconsin, USA. The image captured were from clinic signs, doors, and stairs. It is open source dataset and offered for research, education and academic used [30]. There are three different categories from the dataset that consists of various images which are 754 doors, 702 signs, and 599 stairs. Fig. 3 shows some sample images from each category. The images were captured with variety of view point and intra-class variation with occlusion across each class [18].

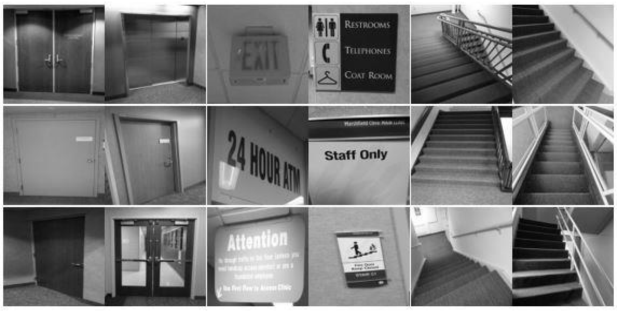

Fig. 3. Sample images from each category in MCIndoor20000 dataset [30].

B. *Experiment Environment*

The experiment environment used for the training and validation of the images are using MatlabR2018a software with DELL Latitude 3580 laptop and Windows 10 Pro for the operating system. The hardware consists of 2.50GHz Intel® Core™ i5-7200U CPU processor and 8 GB of memory.

IV. Results and Discussion

A. *Basic Convolutional Neural Network (CNN)*

The image input size is 250-by-250-by-3, which represents the height, width, and the channel size. The channel size 3 corresponds to the color channel, which are Red, Green, and Blue (RGB) values. Table II shows the accuracy result of basic CNN (Batch normalization and ReLU layers are omitted from the table for brevity). By referring to Table II, Layer 1 indicates a series of layers that are convolutional layer, pooling layer and ReLU layer. Layer 2 means that these series of layers are doubled while Layer 3 is where the number of layers is tripled.

The first convolutional layer uses 3-by-3 as the filter size and 16 as the number of feature maps. A padding of 1 is to ensure that the spatial output size is the same as the input size. Batch normalization layers are used between convolutional layers and nonlinear layer to speed up the network training and lessen the sensitivity to network initialization. ReLU layer is a nonlinear layer followed by batch normalization layer as the activation function. The max-pooling operations are implemented with a stride size of 3 and 3-by-3 filters. Training set of 10, 120, and 300 images per category has been conducted to achieve higher accuracy.

The second layer are executed with the same operation for training set of 10 and 300 images per category, but the highest accuracy achieved is only 92.64%. Another attempt is performed in the third layer for training set of 10 and 300 images per category and 32 as the number of feature maps. The training set of 300 images per category has achieved the highest accuracy which is 97.92%. All training processes use 10 Epoch and 0.001 as the Learning Rate. A slightly smaller or bigger values than 0.001 for the learning rate reduces the accuracy rate.

This can conclude that basic CNN can achieve high accuracy if fed with many training images. This experiment has a total of 2,055 images, with 300 images per category reserved for training images, which is 900 and the remaining is validation images, which is 1,155. The training images are about 44% of the total images. Table II lists the accuracy result of basic CNN.

TABLE II: ACCURACY RESULT OF BASIC CNN

Number of Layer	Training Set / Category	Convolve Layer, Padding	Pooling layer, Stride	Accuracy (%)	Total Time (s)
1	10	3/16, 1	3, 3	55.9	101
	300	3/16, 1	3, 3	92.21	1001
2	10	3/16, 1	3, 3	72.79	104
		3/16, 1	3, 3		
	120	3/16, 1	3, 3	92.28	858
		3/32, 1	3, 3		
	300	3/16, 1	3, 3	92.64	660
		3/16, 1	3, 3		
3	10	3/16, 1	3, 3	81.04	104
		3/16, 1	3, 3		
		3/32, 1	3, 3		
	300	3/16, 1	3, 3	97.92	1630
		3/16, 1	3, 3		
		3/32, 1	3, 3		

B. AlexNet

The input images with size of 227-by-227-by-3 are required by this pre-trained model network. An augmented image datastore is used to automatically resize the training images due to variety image size in the image datastore without specifying any further preprocessing tasks. The initial learning rate is set to 0.0001 as to delay the learning process in the transmitted layers. Epoch is set to 6 since there is no need to train for as many epochs when performing transfer learning. The data is divided into training and validation data sets. Each test used different size of training set. The first test used 540 images and 90 iterations per epoch. The second test used training set of 246 images and 41 iterations per epoch. The third test used 984 training images with 164 iterations per epoch. Table III shows the accuracy result of AlexNet transfer learning.

TABLE III: ACCURACY RESULT OF ALEXNET

Number of Test	Training Set	Epoch	Learning Rate	Accuracy (%)	Total Time (s)
1	540	6	0.0001	99.65	7603
2	246	6	0.0001	99.88	4725
3	984	6	0.0001	100	5787

The tests showed that with bigger size of training set, a higher accuracy can be achieved. The final test consists of 80% of training and 20% of validation images have achieved 100% accuracy with 96 minutes and 27 seconds total training time.

Fig. 4 displays some sample validation images with their predicted labels after classification and validation processes applied by AlexNet.

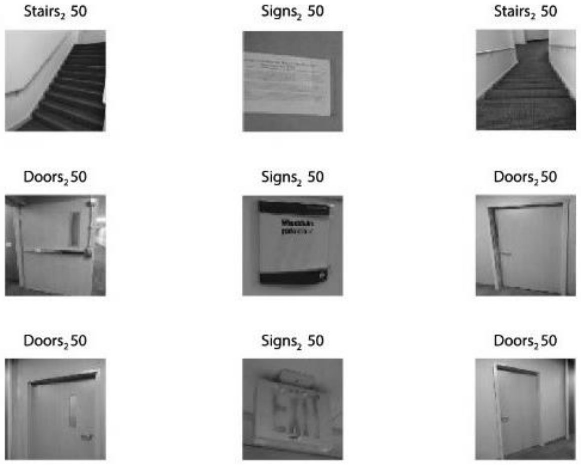

Fig. 4. Sample validation images with their predicted labels.

C. Bag of Features (BoF)

The training set images need to be balanced but the images in the image datastores comprises an unequal number of images for each category. Balancing the images per category is performed by trimming the set with the smallest number of images in a category. Stairs images contain the smallest figure, which is 599, therefore, the image datastores will have 599 images for each category and the total number of images for this experiment is 1,797. Next, the sets are divided into training and validation data with 20% of images from each set for the training data and the remainder 80% for the validation data. Bias result avoided by performing randomize function.

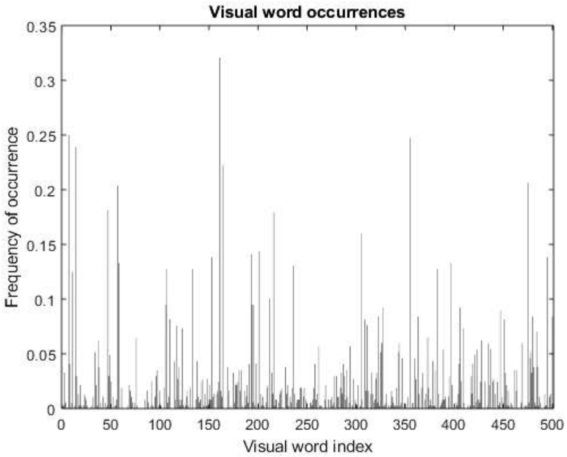

Fig. 5. Histogram of classification into three different categories: Doors, Signs & Stairs.

Extracting SURF features from the selected feature point location which is 1,383,840 features from 360 images. The K-means clustering algorithm is implemented to quantization of the feature space. This will construct 500 visual words vocabulary by reducing the number of 1,107,072 features. Clustering process is completed with 23/100 iterations and processing time is ~7.07 seconds/iteration and converged in

23 iterations. Fig. 5 shows the histogram of the counting of the visual word appearances in an image for this experiment.

The histogram in Fig. 5 forms the base for training a classifier and for the classification of an actual image. Next is to evaluate the classifier performance by using the trained classifier against the training set. Fig. 6 shows the confusion matrix for this test after evaluating the 360 images with average accuracy of 98%.

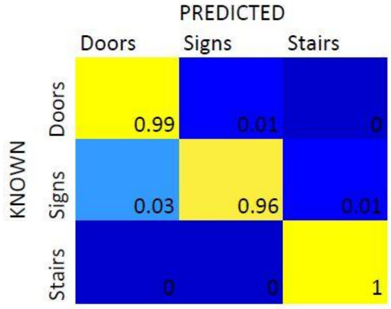

Fig. 6. Confusion matrix for training set.

Then, evaluation of the classifier against the validation set is performed. Fig. 7 shows the confusion matrix for this test after evaluating the 1,437 images with average accuracy of 97%. The high accuracy which is 97% shows that image classification using BoF can differentiate images based on the categories that are door, sign, and stairs for the MCIndoor20000 dataset as good as basic CNN.

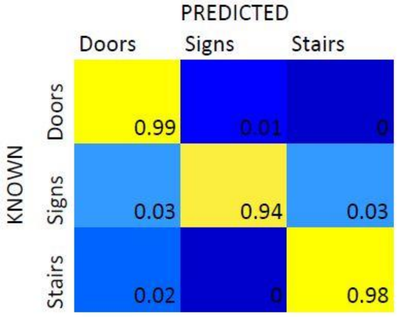

Fig. 7. Confusion matrix for validation set.

V. Conclusion

This paper evaluates the indoor object performance among three techniques namely; AlexNet, basic CNN, and BoF. The results demonstrate that different by using different parameter value able to produce diverse accuracy. Thus, various experiments need to be conducted in order to achieve the desired accuracy. In these experiments, by increasing the size of the training set can improve the accuracy but it will impact the training performance of the classifier where 100% accuracy is achieved with AlexNet but have caused the machine to run about 96 minutes for the 984 of training images. The number of layers also affects the accuracy where the higher number of layers, the longer it needs to achieve the result. BoF produces almost similar accuracy as basic CNN which proves that machine learning is still as good as deep learning. Experimental results indicated that AlexNet achieves 100% accuracy while basic CNN produces 97.92% accuracy and BoF accomplishes 97% accuracy. These results showed that BoF, a machine learning technique, can produce a high accuracy performance as basic CNN, a deep learning technique for image recognition. Future work is to experiment on other features and classifiers for BoF and other types of CNN such as recurrent CNN for object recognition.

Acknowledgment

The authors would like to thank the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia, for supporting this research.

References

[1] S.-S. Shai and B.-D. Shai, *Understanding Machine Learning: From Theory to Algorithms*, New York: Cambridge University Press, 2014.

[2] M. M. Saufi, M. A. Zamanhuri, N. Mohammad, and Z. Ibrahim, "Deep learning for Roman handwritten character recognition," *International Journal of Electrical Engineering and Computer Science*, vol. 12, no. 2, pp. 455-460, 2018.

[3] Z. Ibrahim, N. Sabri, and D. Isa, "Multi-maxpooling convolutional neural network for medicinal herb leaf recognition," in *Proc. the 6th IIAE International Conference on Intelligent Systems and Image Processing*, Shimane, 2018.

[4] N. A. M. Kasim, N. H. A. Rahman, Z. Ibrahim, and N. N. A. Mangshor, "Celebrity face recognition using deep learning," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 12, no. 2, pp. 476-481, 2018.

[5] M. Tounsi, I. Moalla, and A. M. Alimi, "Supervised dictionary learning in BoF framework for Scene Character recognition," presented at 23rd International Conference on Pattern Recognition (ICPR), Cancun, 2016.

[6] M. Anthimopoulos, L. Gianola, L. Scarnato, P. Diem, and S. G. Mougiakakou, "A food recognition system for diabetic patients based on an optimized bag-of-features model," *IEEE J. Biomedical and Health Informatics*, vol. 18, no. 4, pp. 1261-1271, 2014.

[7] A. J. Siddiqui, A. Mammeri, and A. Boukerche, "Real-time vehicle make and model recognition based on a bag of surf features," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 11, pp. 3205-3219, 2016.

[8] Z. Ibrahim, N. Sabri and D. Isa, "Palm oil fresh fruit bunch ripeness grading recognition using convolutional neural network," *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol. 10, no. 3-2, pp. 109-113, 2018.

[9] R. D. Sharma, S. Tripathi, S. K. Sahu, S. Mittal, and A. Anand, "Predicting online doctor ratings from user reviews using convolutional neural networks," *International Journal of Machine Learning and Computing*, vol. 6, no. 2, p. 149, 2016.

[10] A. A. Almisreb, N. Jamil and N. M. Din, "Utilizing AlexNet deep transfer learning for ear recognition," presented at Fourth International Conference on Information Retrieval and Knowledge Management (CAMP), Kota Kinabalu, 2018.

[11] N. Ali, K. B. Bajwa, R. Sablatnig, S. A. Chatzichristofis, Z. Iqbal, M. Rashid, and H. A. Habib, "A novel image retrieval based on visual words integration of SIFT and SURF," *PloS one*, vol. 11, no. 6, p. e0157428, 2016.

[12] R. Azhar, D. Tuwohingide, D. Kamudi, and N. Suciati, "Batik image classification using SIFT Feature extraction, bag of features and support vector machine," *Procedia Computer Science*, vol. 72, pp. 24-30, 2015.

[13] N. Sabri, Z. Ibrahim, M. M. Saad, N. N. A. Mangshor, and N. Jamil, "Human detection in video surveillance using texture features," presented at 6th IEEE International Conference on Control System, Computing and Engineering (ICCSCE), Penang, 2016.

[14] L. Liu, L. Shao, and P. Rockett, "Human action recognition based on boosted feature selection and naive Bayes nearest-neighbor classification," *Signal Processing*, vol. 93, no. 6, pp. 1521-1530, 2013.

[15] A. Adebowale, S. A. Idowu, and A. Amarachi, "Comparative study of selected data mining algorithms used for intrusion detection," *International Journal of Soft Computing and Engineering (IJSCE)*, vol. 3, no. 3, pp. 237-241, 2013.

[16] J. Park, B. Ko, J. Y. Nam, and S. Kwak, "Wildfire smoke detection using spatiotemporal bag-of-features of smoke," presented at the 2013 IEEE Workshop on Applications of Computer Vision (WACV), Washington, 2013.

[17] W. Chen, T. Qu, Y. Zhou, K. Weng, G. Wang and G. Fu, "Door recognition and deep learning algorithm for visual based robot navigation," presented at the 2014 IEEE International Conference on Robotics and Biomimetics, Bali, 2014.

[18] F. S. Bashiri, E. LaRose, P. Peissig, and A. P. Tafti, "MCIndoor20000: a fully-labeled image dataset to advance indoor objects detection," *Data in Brief*, vol. 17, pp. 71-75, 2018.

[19] T. Janani and A. Ramanan, "Feature fusion for efficient object classification using deep and shallow learning," *International Journal of Machine Learning and Computing*, vol. 7, no. 5, pp. 123-127, 2017.

[20] P. Pawara, E. Okafor, O. Surinta, L. Schomaker, and M. Wiering, "Comparing local descriptors and bags of visual words to deep convolutional neural networks for plant recognition," presented at 6th International Conference on Pattern Recognition Applications and Methods, Porto, 2017.

[21] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, "Land use classification in remote sensing images by convolutional neural networks," *arXiv* preprint arXiv:1508.00092., 2015.

[22] S. Sladojevic, M. Arsenovic, A. Anderla, D. Culibrk, and D. Stefanovic, "Deep neural networks based recognition of plant diseases by leaf image classification," *Computational Intelligence and Neuroscience*, no. 3289801, pp. 1-11, 2016.

[23] J. F. Couchot, R. Couturier, C. Guyeux, and M. Salomon, "Steganalysis via a convolutional neural network using large convolution filters for embedding process with same stego key," *arXiv* preprint arXiv:1605.07946., 2016.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, pp. 1097-1105, 2012.

[25] M. Elleuch, A. M. Alimi, and M. Kherallah, "Enhancement of deep architecture using Dropout/DropConnect techniques applied for AHR system," presented at 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, 2018.

[26] C. F. Tsai, "Bag-of-words representation in image annotation: A review," *ISRN Artificial Intelligence*, no. 376804, pp. 1-19, 2012.

[27] K. Ahmad, R. Khan, N. Ahmad, and J. Khan, "Evaluation of SIFT and SURF using bag of words model on a very large dataset," *Sindh University Research Journal (Science Series)*, vol. 45, no. 3, pp. 492-495, 2013.

[28] B. Zhu, L. Yang, X. Wu, and T. Guo, "Automatic recognition of books based on machine learning," presented at 2015 3rd International Symposium on Computational and Business Intelligence (ISCBI), Bali, 2015.

[29] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *International Journal of Computer Vision*, vol. 73, no. 2, p. 213–238, 2007.

[30] F. S. Bashiri, E. LaRose, P. Peissig and A. P. Tafti. (January 9, 2018). GitHub - bircatmcri/MCIndoor20000. [Online]. Available: https://github.com/bircatmcri/MCIndoor20000

Srie Azrina Zulkeflie completed her bachelor degree in computer science (software engineering) at University Malaya, Kuala Lumpur, Malaysia. Currently, she is pursuing her master's degree in computer science (web technology) at Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia. Her interests are Geographical Information Science (GIS), programming. She is currently active in indoor positioning and indoor navigation research. She is a professional member of Institution of Geospatial and Remote Sensing Malaysia (IGRSM).

Fatin Amira Fammy completed her bachelor of technology degree in data communication and networking at Universiti Teknologi MARA (UiTM), Jasin, Melaka, Selangor. She is currently pursuing her master degree in web technology at Universiti Teknologi MARA (UiTM), Shah Alam, Selangor, Malaysia. Her interests include machine learning and data analysis.

Zaidah Ibrahim is an associate professor at the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia. She is an active member of Digital Image, Audio and Speech Technology (DIAST) research group where she has presented papers in areas related to computer vision at local and international conferences and published in journals. She has also been awarded with a few research grants. Her current interest is the application of machine learning and deep learning in object detection and recognition.

Nurbaity Sabri is a lecturer at the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Jasin, Melaka, Malaysia. She teaches programming languages and image processing. She is a member of Digital Image, Audio and Speech Technology (DIAST) research group and currently participating in various researches related to image processing. She has published papers and co-authored in international conferences and journals. Her research interests include image processing, computer vision and pattern recognition.