ZANCO Journal of Pure and Applied Sciences
The official scientific journal of Salahaddin University-Erbil
https://zancojournals.su.edu.krd/index.php/JPAS

ISSN (print):2218-0230**, ISSN (online):** 2412-3986**, DOI: http://dx.doi.org/10.21271/zjpas**

RESEARCH PAPER

Indoor Place Recognition and Localization Using Histogram of Oriented Gradient with Deep Learning

Zina Khaleel Jalal [1],Moayad Y. Potrus[2], Abbas M. Ali[3]

Department of Software and Informatics Engineering,College of Engineering, Salahaddin University-Erbil,Kurdistan Region, Iraq

A B S T R A C T:

Indoor place recognition is a crucial and challenging field of computer science. It is widely used in robotics and computer vision for various applications. The challenges in indoor place recognition comes from the fact that recognizing localized places like office, corridor, and others may fall under various environmental effects of weather, illumination and others. In this paper, an indoor place recognition and localization system is proposed. The system utilizes the great recognition capabilities of Convolutional Neural Network (CNN) and AlexNet with the use of feature image for training. The feature images are constructed using Histogram of Gradient (HOG). The main contribution of this work is the use of 2D feature constructed image from HOG instead of the scene image used with CNN. The proposed system was compared to other previous systems, in which, it achieved better recognition accuracy when tested on COLD and IDOL standard indoor image datasets.

KEY WORDS: Place recognition, Localization, CNN, AlexNet, SIFT, HOG
DOI: http://dx.doi.org/10.21271/ZJPAS.32.1.3
ZJPAS (2020) , 32(1);19-30 .

1.INTRODUCTION :

Place recognition can be described as a technique that allows robots to determine if a captured image of a place has been visited before or not (Bai et al., 2018). Actually, robots must be capable of working in entirely various places, under many different environmental conditions (Mancini et al., 2018).

To realize long term autonomy and localization, specific problems in the area of changing environments must be treated. These environmental changes may include illumination changes, weather and point of view, which impact the accuracy of place recognition(Kumar et al., 2017).

* **Corresponding Author:**
Zina Khaleel Jalal
E-mail: zenaalassaff@gmail.com
Article History:
Received: 27/07/2019
Accepted: 29/09/2019
Published: 25/02 /2020

Despite of many researches proposed for the reinforcement of place recognition, enhancing reliability and accuracy of place recognition, but still it's a challenging issue that needs an optimum solution (Kumar et al., 2017). Place recognition techniques traditionally depend on representing images visual content by utilizing local features, like SURF (Bay et al., 2006), SIFT (Lowe, 2004) or by means of "Bag of visual Words (Bovw)" model(Kenshimov et al., 2017). In computer vision, after the remarkable successfulness of deep learning, the focus of place recognition studies has lately changed from using conventional handcrafted features like SURF or SIFT to more generic features based on deep learning (Chen et al., 2017b).

Deep learning can be described as a particular branch of machine learning that utilizes multiple layers that consist of nonlinear transformation units (Chollet, 2018) one of the networks used for

deep learning is Convolutional Neural Network. CNN is categorized as a deep learning model (Li et al., 2011). CNNs involve a large number of layers as Compared to original neural networks (Tobías et al., 2016). Convolutional neural networks are attaining significance in various tasks (Lopez-Antequera et al., 2017) like affordance prediction (Porzi et al., 2016), object classification (He et al., 2016) and depth estimation (Xu et al., 2017).

A basic convolutional neural network and AlexNet are presented in this work to perform place recognition under several illumination conditions such (cloudy, sunny and night) for training and testing both models. Two commonly available datasets, COLD and IDOL, are utilized for this purpose. HOG and SIFT were utilized with both networks for calculating the accuracy of recognition. The proposed modification and contribution of the work is achieved by transforming the Hog and Sift feature vector into 2D vector for CNN training. The organization of the research paper is in this order: Section 2 represents the related works. Section 3 introduces work steps, feature extractors and the networks that have been utilized. Section 4 introduces the results and comparison results that have been achieved in this research work. Finally, Section 5 illustrates our conclusion.

2. Related works

Place recognition has been focused on by many researchers and a huge number of works have done in this field. FAB-MAP by (Cummins and Newman, 2008) was one of the earliest works which proposed the utilization of Bag of words in place recognition. In their work, depending on the appearance of every position, suggested the probabilistic way of place recognition.

(Park et al., 2018) Presented a light-weight visual place recognition which is depending on CNN. The presented architecture is particularly designed for mobile robots which supplied with embedded systems. The architecture has a fewer filter and five convolutional layers. According to the computational time and accuracy in custom and KTHIDOL2 datasets, the results outperformed the traditional algorithms based on CNN approaches.

Researchers in (Lopez-Antequera et al., 2017) proved that the task of place recognition can be solved better by training a network discriminatively. They trained CNN under appearance changes like seasons, weather, point of view and time of day. For training CNN, they utilized triplet-based learning schema to place images in a low dimensional space. Such that, where small Euclidean distances are delegating of place sameness and conditions are demanding. They claimed that their presented network outperformed all general methods.

A new approach was suggested by (Kenshimov et al., 2017) which addresses the issue of the cross-season place recognition. It has the ability to raise the robustness of cross-season place recognition. It creates an image descriptor of lower dimensionality through deleting the activation of filters which correlates to environmental variations. Then, to extract the entire output of an intermediate layer of CNN.

For treating domain generalization in the state of semantic place classification, a new version of a deep learning model introduced by researchers in (Mancini et al., 2018). They built CNN architecture with added layers of weighted form of batch normalization. The architectures were AlexNet and ResNet. They tested the architecture on COLD dataset with various illumination conditions like sunny, night and cloudy. Their experimentations proved that using the new WBN layers on visual place recognition benchmark, achieved a better accuracy of place categorization.

3. Materials and methods:

Indoor Place recognition using extracting features then analyzing and recognizing them is a common method in the field. The proposed system uses CNN and AlexNet with Hog features. The whole process has been illustrated in Figure 1. The following subsection detail each of the proposed framework steps.

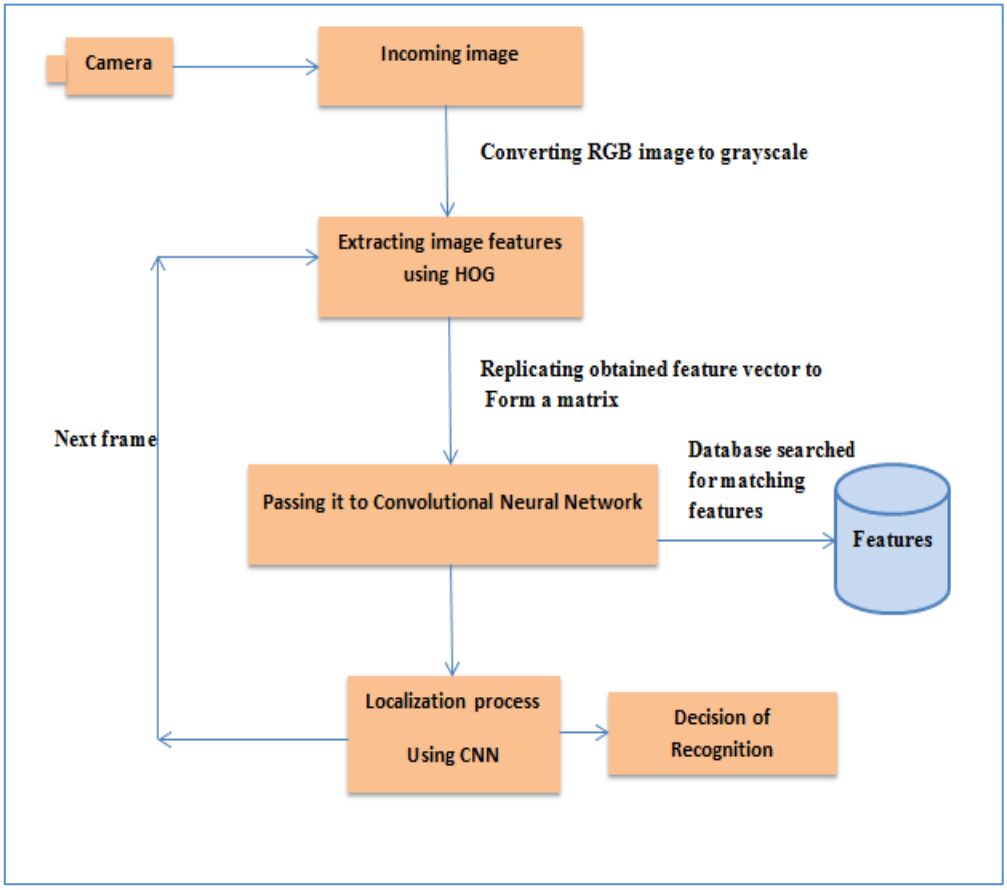

Figure 1: The whole process of proposed work

3.1.Pre-processing:

Images were imported from the COLD and IDOL datasets which are RGB images of size 640x480 pixels and 309x240 pixels respectively. Then these images were converted to grayscale images for next step of feature extraction.

3.2. Feature extraction:

In this section, feature extraction methods which have been utilized in this work such as HOG (Dalal and Triggs, 2005) and SIFT (Ledwich and Williams, 2004) with soft assignment of BOW, are discussed.

The following sections explain these methods.

3.2.1. Histogram of Oriented Gradient (HOG):

HOG is among the widespread features suggested by (Dalal and Triggs, 2005). The fundamental concept beyond HOG features is that appearance and local object shape inside an image can be described by the distribution of edge trends or intensity gradients. HOG partitions the image into small joint regions, named cells, and for every cell, it combines a histogram of gradient trends for the pixels inside the cell. Every pixel within the cell throws a weighted vote for an orientation-based histogram channel based on the values found in the gradient computation. The histogram channels are equally publishing over 0 to 180 degrees. The feature vector was represented by the integration of these cell histograms (Ren and Li, 2014). After generating feature vectors from HOG which has been produced nx1 feature vector, then the feature vector is horizontally stacked a column vector n times to generate a square matrix to be ready for input to both models. Figure 2 illustrates our contribution to HOG.

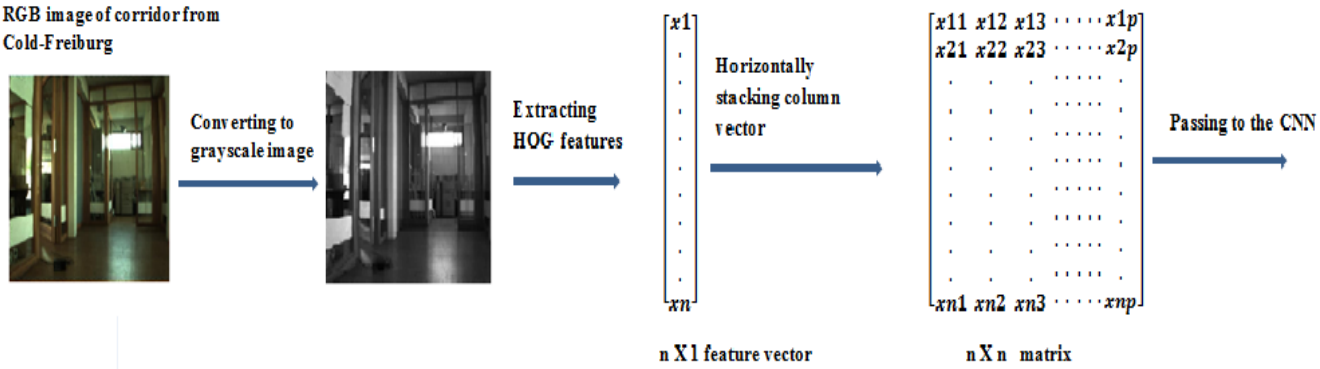

Figure 2: Process of converting HOG feature vector to a square matrix

3.2.2. Scale invariant feature transform (SIFT):

SIFT can be defined as one of the widely utilized local visual descriptors. It includes two steps in which, the first step, is about feature detection of images, the second step describes the extracted features (Sykora et al., 2014). SIFT detector is strong and stable to rotations, scaling, translation and partly stable to illumination variations and viewpoint of the camera (Mansourian et al., 2015). There are four stages in the algorithm. First, the images are checked under different octaves and scales for insulating image points that are different from their environments these points are defined as extreme points. Thereafter, points badly placed on an edge plus points which have low contrast have been removed (Mansourian et al., 2015). After this depending on local image properties a stable orientation is appointed to the key points. The keypoint descriptor usually utilizes a set of 16 histograms. This is placed in a 4x4 grid and each grid has 8 orientation bins. There is one for every of the major compass trends and for every central-points of these trends. This generates a feature vector which includes 128 items. Figure 3 shows generating keypoint descriptor (Lowe, 2004).

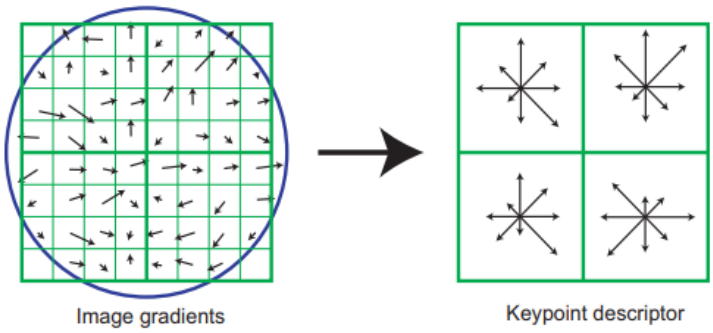

Figure 3: Generation of keypoint descriptor (Lowe, 2004)

3.2.3. Bag of Words (BOW):

BOW is defined as one of the most common representation ways for object classification. The basic concept behind it is to quantize every extracted key point into one of the visual words, and after that, every image will be represented by a histogram of visual words. For creating visual words, the mostly utilized clustering algorithm is (K-means) as shown in Figure 4 (Zhang et al., 2010).

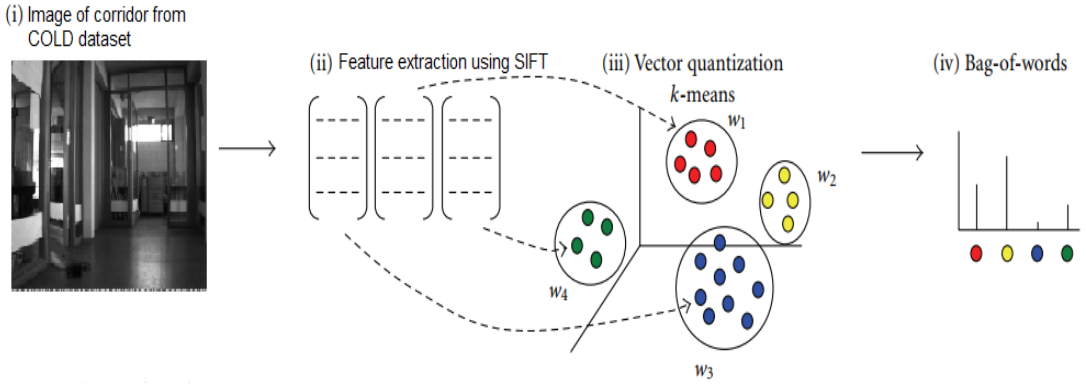

Figure 4: Steps for constructing bag of word

3.3. Convolutional Neural Network (CNN):

Feature vectors which are obtained from feature extractors are converted to a square matrix then passed as input to the CNN. After the localization process by CNN, the next frames will be read from the live images thereafter returning back to the feature extraction step to compute features for new frames. In this work, for training a CNN from scratch, 15 layers are used. These layers are 1 input layer, 3 convolutional layers with the filter size of 3x3, 2 max pooling layers, 3 batch normalization layers, 3 rectified liner unit layers, 1 fully connected layer, 1 softmax layer and 1 classification layer.

The following are the description of the layers:

3.3.1. Input layer:

It is the first layer which takes images and resizes them for passing it into the next layers (Sharma et al., 2018).

3.3.2. Convolution layers:

Convolution layers are distinguished through filter values. For each layer, there are multiple convolutions with a stable size also every kernel with a stable stride applied over the complete image (Tobías et al., 2016). Here the features of images will be found and then passed into the pooling layer (Sharma et al., 2018) Equation(1) represents the convolution operation (Chen et al., 2017a):

$$y^j = \max\left(0, b^j + \sum_i k^{ij} * x^i\right) \qquad (1)$$

y^j is the j-th output map and x^i is the i-th input map, k^{ij} represents the convolution kernel between j-th output map and the i-th input map also * represents the convolution operation.

3.3.3. Max Pooling Layer:

This layer takes the big sized images and shrinks them down while keeping the most significant information in them. From every window, it retains the maximum value; it keeps the finest fits of every feature in the window (Sharma et al., 2018). The operation of max pooling is illustrated in Equation (2) (Chen et al., 2017a).

$$y^i_{j,k} = \max_{0 \le m,n \le r}\left(x^i_{jxr+m,kxr+n}\right) \qquad (2)$$

Every activation $_{yj,k}^{i}$ in the i-th pooling map $_{y}^{i}$ pools above $r\ by\ r$ non overlapping region in the i-th input map $_{x}^{i}$.

3.3.4. Rectified Linear Units Layers (ReLU):

Each negative number of the pooling layer will be changed with zero. This layer is utilized after the convolution layer, which aids the CNN to remain mathematically steady (Sharma et al., 2018). It is represented by Equation (3) below (Tobías et al., 2016):

$$F(x) = max\,(0, x) \quad (3)$$

3.3.5. Batch normalization layer:

It is another layer that can be placed in the architecture of the model, such as a convolutional or fully connected layer. It supplies an introduction for input feed-forwarding and calculating gradients with regard to the parameters (Schilling, 2016).

3.3.6. Softmax layer:

Generally, deep learning provides a solution for classification problem, by utilizing a function of softmax as their classification function (final layer). The function of softmax identifies the discrete probability P allocation for K classes. This can be indicated by $\sum_{k=1}^{k} pk$ (Yosinski et al., 2014).

If x be as the activation, and θ be as its weight parameters at the softmax layer, then o can be represented as input to the softmax layer,

$$o = \sum_{i}^{n-1} \theta ixi \quad (4)$$

Consequently

$$pk = \frac{\exp(o_k)}{\sum_{k=0}^{n=1} \exp(o_k)} \quad (5)$$

Hence, the predicted class would be $\hat{y}$

$$\hat{y} = \underset{i \in 1 \ldots N}{\arg\max}\ pi \quad (6)$$

3.4. AlexNet:

AlexNet is a vastly utilized CNN architecture (Tobías et al., 2016). For this work, basic AlexNet architecture is used which is pre-trained on ImageNet (Deng et al., 2009). The network includes 3 fully-connected and the collection of 5 convolutional layers. The latest FC layer is linked to the softmax layer resulting in 1000 class. For complete experiments on AlexNet in this work, the inputs which are obtained from replicating feature vectors are rescaled to 227x227x3.This will fit the feature vector into the network. In addition, the latest fully connected layer are fine-tuned to output 5 classes instead of 1000 as this paper has five classes used from each of COLD and IDOL datasets. Fine-tuning (Yosinski et al., 2014) means utilizing a formerly trained network as the starting of the training step, for a given task and training datasets (Tobías et al., 2016). Figure 5 presents the architecture of pre-trained AlexNet.

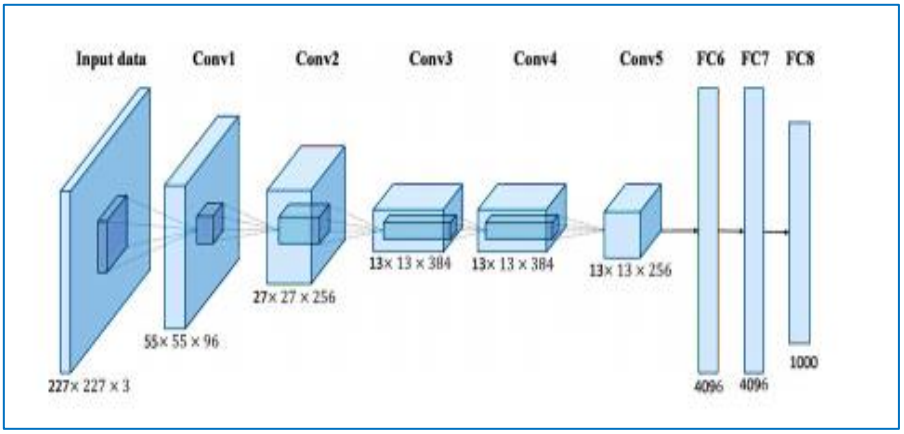

Figure 5: AlexNet architecture (Han et al., 2017)

4. Results and Discussion

In the proposed work, for calculating recognition accuracy, CNN and AlexNet have been used with two feature extractor methods HOG and SIFT. They both were applied to two common datasets COLD and IDOL.

COLD stands for Cosy Localization Database which includes three autonomously gathered sub-datasets (COLD-Freiburg, COLD-Ljubljana and COLD Saarbrucken) collected from three distinct interior laboratory environments using different robots. The COLD database is a typical testbed for evaluating the robustness of place recognition algorithms and localization. It is captured according to the dynamic and categorical changes established by human activity and illumination changes (Pronobis and Caputo, 2009). Figure 6 presents some images from the COLD database.

IDOL is an abbreviation which points to an Image Database for Robot Localization. The database includes 24 image series obtained by utilizing two moving robot platforms. The acquisition was carried out inside an interior laboratory environment which contains five rooms of various employments (corridor, two-person office, one-person office, printer area, and kitchen). Moreover, it was acquired under different illumination changes such as night, sunny and cloudy (Luo et al., 2006). Figure 7 presents some images from IDOL database.

For the COLD dataset when CNN utilized with SIFT gets higher recognition accuracy in the COLD night images which is 94.77%. Also, AlexNet with SIFT achieves higher accuracy in the COLD night images with accuracy of 92.47%. The results show that CNN has better accuracy than AlexNet as shown in Table 1.

For improving the accuracy of recognition, HOG has been used with both CNN and AlexNet. In the COLD dataset, CNN with HOG achieved higher accuracy in COLD cloudy images of about 98.36%, and for AlexNet by using HOG features achieves more accurate results in COLD night images which are 97.07%. These results are shown in Table 2. The results show a noticeable improvement in the recognition accuracy by using HOG rather than SIFT. Furthermore, for the IDOL dataset, CNN with HOG achieves higher accuracy for Idol night images of 97.31%. Also, for AlexNet with HOG in the same dataset gets higher accuracy for Idol night images of 95.03%. The results are presented in Table 3.

The achieved results of AlexNet by using HOG for the COLD dataset is compared with the work of (Mancini et al., 2018) for place recognition. In the work of Mancini et al. a batch normalization (BN) layer was added after each fully connected layer of Alex Net, but in the proposed work base AlexNet has been used and achieves higher accuracy for COLD night images it achieved better results as average of accuracy which is 96.13% compared to Mancini et al average accuracy of 94.6%. These are for the COLD-Freiburg group as shown in Figure 8.

Moreover, the proposed work added BN layer after each convolutional layer of base AlexNet with HOG features. The proposed setup was compared with the work of Mancini et al, which added the BN layer after each fully connected layer. The proposed AlexNet using BN layer gave a better accuracy for both COLD sunny and COLD night images and obtained better results with average of accuracy of 97.38% compared to Mancini work of 94.6% average of accuracy. As shown in Figure 9. These implementations are only for the COLD Freiburg group. The proposed CNN+HOG was compared with the work of (Mancini et al., 2017) which integrates CNN with Naive Bayes Nearest Neighbor (NBNN) model in the COLD-Freiburg dataset. The results showed that the proposed CNN achieves higher average accuracy than the work of Mancini et al., 2017. This is illustrated in Table 4.Finally, the proposed Alexnet+HOG and CNN+HOG were compared with the (ResNet + BN) and (ResNet+ Weighted BN (WBN)) by Mancini et al., 2018 applied to COLD-Freiburg dataset. The proposed work outperforms ResNet network which presented in Table 5.

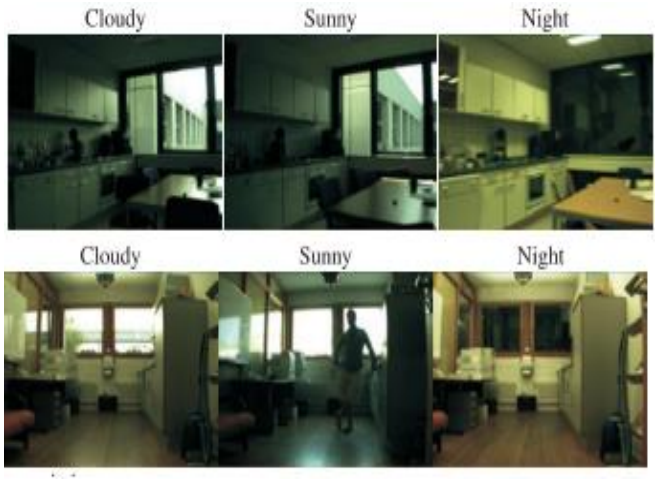

Figure 6: Images under different illumination from Cold dataset (Pronobis and Caputo, 2009)

Figure 7: Sample images from Idol dataset with various Illuminations (Luo et al., 2006)

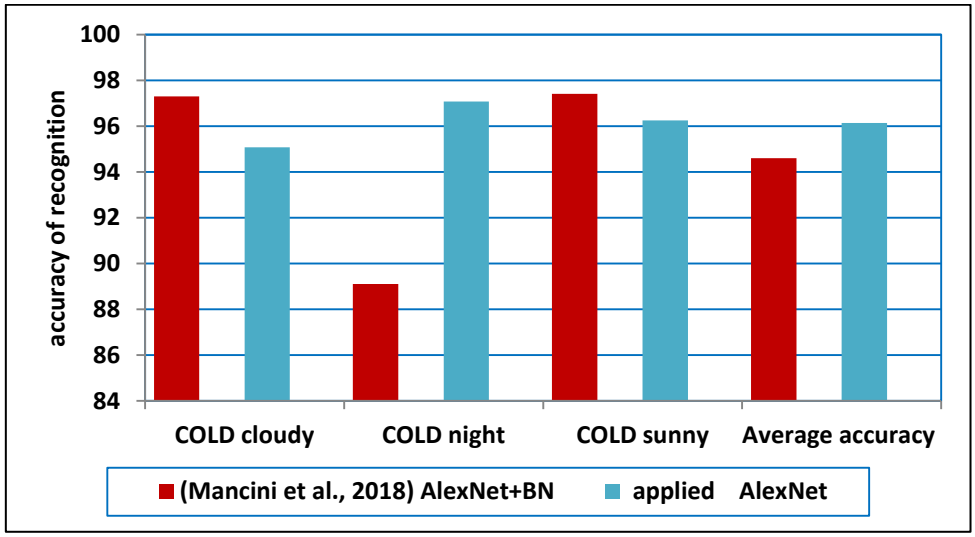

Figure 8: Comparing results for applied (AlexNet by utilizing HOG) and (AlexNet with BN layer) in (Mancini et al., 2018) for Cold-Freiburg.

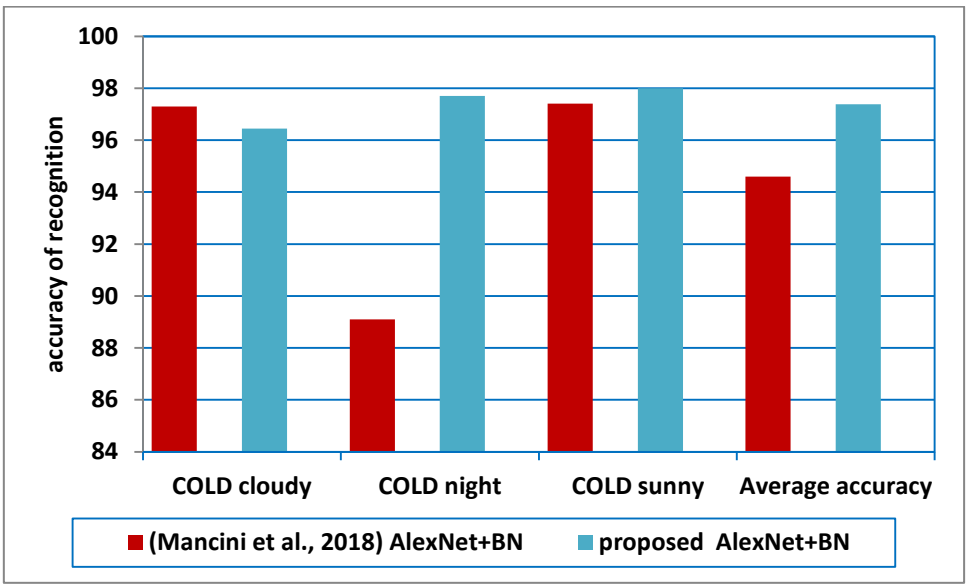

Figure 9: Comparing results of proposed (AlexNet with BN layer) and (AlexNet with BN layer) in (Mancini et al., 2018) for Cold-Freiburg.

Table 1: Recognition accuracy of CNN and AlexNet by utilizing SIFT in COLD dataset

SIFT with model	COLD cloudy	COLD sunny	COLD night
SIFT + CNN	88.49%	92.25%	94.77%
SIFT + AlexNet	86.58%	86.75%	92.47%

Table 2: Recognition accuracy of CNN and AlexNet by utilizing HOG in COLD dataset

HOG with model	COLD cloudy	COLD sunny	COLD night
HOG + CNN	98.36%	98.25%	97.49%
HOG + AlexNet	95.07%	96.25%	97.07%

Table 3: Recognition accuracy of CNN and AlexNet by utilizing HOG in IDOL dataset

HOG with model	IDOL cloudy	IDOL sunny	IDOL night
HOG + CNN	96.51%	97.09%	97.31%
HOG + AlexNet	93.68%	92.84%	95.03%

Table 4: Results of comparison of proposed (CNN + HOG) with (CNN + NBNN)

Method	Average accuracy on COLD-Freiburg
CNN with NBNN(Mancini et al., 2017)	95.2
Proposed CNN with HOG	**98.03**

Table 5: Results of comparison of proposed CNN, AlexNet with ResNet (Mancini et al., 2018)

Method	Average accuracy on COLD-Freiburg
ResNet + BN (Mancini et al., 2018)	90.2
ResNet + WBN(Mancini et al., 2018)	91.33
Proposed (AlexNet + BN) with HOG	**97.38**
Proposed CNN with HOG	**98.03**

5. Conclusion

In this research paper, CNN and AlexNet have been used for indoor place recognition and localization. This work attempted to find out the accuracy of recognition of two models applied to Further, when comparing AlexNet result with the work of Mancini et al., 2018 it demonstrated its superior performance in getting better results in terms of accuracy average. Moreover, the proposed CNN+HOG is compared with CNN+NBNN Mancini et al.,2017 in COLD-Freiburg dataset that shows proposed CNN two common datasets (COLD and IDOL) under different illumination conditions such as cloudy, night and sunny. The results on these two datasets demonstrated that the proposed CNN with HOG features significantly outperforms the AlexNet. outperforms CNN+NBNN. Finally, CNN+HOG and proposed (AlexNet+BN with HOG) are compared with the CNN architecture (ResNet with (BN and WBN)) in Mancini et al., 2018 which showed the proposed networks achieve higher accuracy average for recognition than of ResNet.

References

BAI, D., WANG, C., ZHANG, B., YI, X. & YANG, X. 2018. Sequence searching with CNN features for robust and fast visual place recognition. Computers & Graphics, 70, 270-280.

BAY, H., TUYTELAARS, T. & VAN GOOL, L.2006. Surf: Speeded up robust features. European conference on computer vision.Springer, 404-417.

CHEN, Z., JACOBSON, A., SÜNDERHAUF, N., UPCROFT, B., LIU, L., SHEN, C., REID, I. & MILFORD, M.2017. Deep learning features at scale for visual place recognition.IEEE International Conference on Robotics and Automation (ICRA). IEEE, 3223-3230.

CHEN, Z., MAFFRA, F., SA, I. & CHLI, M.2017. Only look once, mining distinctive landmarks from convnet for visual place recognition.IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 9-16.

CHOLLET, F. 2018. Deep Learning mit Python und Keras: Das Praxis-Handbuch vom Entwickler der Keras-Bibliothek, MITP-Verlags GmbH & Co. KG.

CUMMINS, M. & NEWMAN, P. 2008. FAB-MAP: Probabilistic localization and mapping in the space of appearance. The International Journal of Robotics Research, 27, 647-665.

DALAL, N. & TRIGGS, B.2005. Histograms of oriented gradients for human detection.

DENG, J., DONG, W., SOCHER, R., LI, L.-J., LI, K. & FEI-FEI, L.2009. Imagenet: A large-scale hierarchical image database. IEEE conference on computer vision and pattern recognition. Ieee, 248-255.

HAN, X., ZHONG, Y., CAO, L. & ZHANG, L. 2017. Pre-trained alexnet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification. Remote Sensing, 9, 848.

HE, K., ZHANG, X., REN, S. & SUN, J.2016. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition.770-778.

KENSHIMOV, C., BAMPIS, L., AMIRGALIYEV, B., ARSLANOV, M. & GASTERATOS, A. 2017. Deep learning features exception for cross-season visual place recognition. Pattern Recognition Letters, 100, 124-130.

KUMAR, D., NEHER, H., DAS, A., CLAUSI, D. A. & WASLANDER, S. L.2017. Condition and viewpoint invariant omni-directional place recognition using cnn. 14th Conference on Computer and Robot Vision (CRV). IEEE, 32-39.

LEDWICH, L. & WILLIAMS, S. 2004. Reduced SIFT features for image retrieval and indoor localisation. Australian conference on robotics and automation. Citeseer, 3.

LI, P., LEE, S.-H. & HSU, H.-Y. 2011. Review on fruit harvesting method for potential use of automatic fruit harvesting systems. Procedia Engineering, 23, 351-366.

LOPEZ-ANTEQUERA, M., GOMEZ-OJEDA, R., PETKOV, N. & GONZALEZ-JIMENEZ, J. 2017. Appearance-invariant place recognition by discriminatively training a convolutional neural network. Pattern Recognition Letters, 92, 89-95.

LOWE, D. G. 2004. Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60, 91-110.

LUO, J., PRONOBIS, A., CAPUTO, B. & JENSFELT, P. 2006. The kth-idol2 database. KTH, CAS/CVAP, Tech. Rep, 304.

MANCINI, M., BULÒ, S. R., CAPUTO, B. & RICCI, E. 2018. Robust place categorization with deep domain generalization. IEEE Robotics and Automation Letters, 3, 2093-2100.

MANCINI, M., BULÒ, S. R., RICCI, E. & CAPUTO, B. 2017. Learning deep NBNN representations for robust place categorization. IEEE Robotics and Automation Letters, 2, 1794-1801.

MANSOURIAN, L., ABDULLAH, M. T., ABDULLAH, L. N. & AZMAN, A. 2015. Evaluating classification strategies in bag of sift feature method for animal recognition. Research Journal of Applied Sciences, Engineering and Technology, 10, 1266-1272.

PARK, C., JANG, J., ZHANG, L. & JUNG, J.-I. 2018. Light-weight visual place recognition using convolutional neural network for mobile robots. IEEE International Conference on Consumer Electronics (ICCE). IEEE, 1-4.

PORZI, L., BULO, S. R., PENATE-SANCHEZ, A., RICCI, E. & MORENO-NOGUER, F. 2016. Learning depth-aware deep representations for robotic perception. IEEE Robotics and Automation Letters, 2, 468-475.

PRONOBIS, A. & CAPUTO, B. 2009. COLD: The CoSy localization database. The International Journal of Robotics Research, 28, 588-594.

REN, H. & LI, Z.-N.2014.Object detection using edge histogram of oriented gradient. IEEE International Conference on Image Processing (ICIP). IEEE, 4057-4061.

SCHILLING, F. 2016. The effect of batch normalization on deep convolutional neural networks.

SHARMA, N., JAIN, V. & MISHRA, A. 2018. An analysis of convolutional neural networks for image classification. Procedia computer science, 132, 377-384.

SYKORA, P., KAMENCAY, P. & HUDEC, R. 2014. Comparison of SIFT and SURF methods for use on hand gesture recognition based on depth map. AASRI Procedia, 9, 19-24.

TOBÍAS, L., DUCOURNAU, A., ROUSSEAU, F., MERCIER, G. & FABLET, R. 2016 Convolutional Neural Networks for object recognition on mobile devices: A case study. 23rd International Conference on Pattern Recognition (ICPR). IEEE, 3530-3535.

XU, D., RICCI, E., OUYANG, W., WANG, X. & SEBE, N. 2017. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 5354-5362.

YOSINSKI, J., CLUNE, J., BENGIO, Y. & LIPSON, H. 2014. How transferable are features in deep neural networks? Advances in neural information processing systems. 3320-3328.

ZHANG, Y., JIN, R. & ZHOU, Z.-H. 2010. Understanding bag-of-words model: a statistical framework. International Journal of Machine Learning and Cybernetics, 1, 43-52.