Intelligent Scene Recognition and understanding Basing on Deep Learning Models and Image Databases

Fawaz Albalawi[1*], Yousef Alanazi[1†], Hamad Alyami[1‡], Wassim Messoudi[1§] and Tareq Alhmiedat [1,2¶]

[1] Faculty of Computers and Information Technology, University of Tabuk, Tabuk, Saudi Arabia
[2] Industrial Innovation & Robotics Center, University of Tabuk, Tabuk, Saudi Arabia

Summary
Nowadays, artificial intelligence is currently used in several areas, including industrial, business, and robotics. This research focuses within this context, especially on the intelligent scene recognition, which is considered as challenging task that can provides mobile robots with a higher semantic awareness of their surroundings. In this paper, we present an overview of an intelligent scene recognition process based on several deep learning models in- cluding ResNet50, VGGNet, DenseNet, etc. These models require a training phase based on image databases. In our study, we use existing image databases describing indoor environment such as Places, Scene15, MIT Indoor67, ImageNet, etc. The goal of our research is to determine the suitable image- database and deep learning model for building the scene recog- nition system of indoor environment.

Keywords:
Scene recognition; scene understanding; machine learning; Deep learning; Intelligent image understanding

1. Introduction

Mobile robotics is approaching a degree of maturity thatis starting to allow robots to go out of research labs. Despite advancements, modern robots still have little understanding of their surroundings. Most robots, for example, still use low-level maps to describe their surroundings,which are typically limited to information about occupied and unoccupied locations, low- level visual landmarks, or specific structural limitations. Thereis an urgent need to provide mobile robots with a higher se- mantic awareness of their surroundings to raise the complexityof the jobs they can accomplish in natural situations.

One of the most promising sensor modalities for bridging the semantic gap in today's mobile robots looks to be vision. Most seeing strength and adaptability are evident indicationsof the benefits of a good visual system. For indoor robotic platforms that must interact closely with humans, the abilityto traverse an environment and comprehend the location and type of all objects within it is critical. Furthermore, a new promising paradigm for building strong seeing robots has recently emerged from the utilization of both computer vision and machine learning techniques.

Manuscript received June 5, 2022
Manuscript revised June 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.6.61

This research focuses within this context, especially onthe intelligent scene understanding, which is considered aschallenging task that can provides mobile robots with a higher semantic awareness of their surroundings. In fact, the recent progress of machine learning techniques like deep neural networks, and the emergence of large scale image databases such as ImageNet and Places, has proved the possibility of building efficient intelligent scene understanding system.

The goal of our research is to determine the suitable image-database and deep learning model for building the scene recognition system of indoor environment. To achieve this goal, firstly, we present an overview of an intelligentscene recognition process based on several deep learning models including ResNet50[1], VGG16 [2], DenseNet [3], etc.Secondly, we investigate existing image databases describing indoor environment such as Places365 [4] , Scene15 [5], MIT Indoor67 [6], and ImageNet [7], etc. Finally, we train deep learning models with database images of indoor environment for determining the efficient DL-model / DB Image. These results can help to select the efficient model that we will adoptfor robot navigation.

The rest of the paper is structured as follows: first, we present an overview of the scene recognition and understandingprocess in computer vision. Next, we focus on deep learning based models applied on scene understanding. And then, we presents proposed methodology and experimental results.

2. Scene recognition and understanding overview

2.1 Introduction

Scene understanding is a difficult and crucial problem in computer vision. Images are visual in nature, yet the visual information can take many forms, including shape, edges, texture, and color. Object detection's major goal is to determine what items are there in an image and where they are all placed. Understanding a scene incorporates relevant information at various levels. Humans are most intuitive and natural when it comes to interacting with

various items. Scene understanding, as opposed to object recognition, identifies the goal of objects as well as the distribution of targets in a scene. Scene compre- hension has a substantial impact on computer vision's ability to observe, evaluate, and interpret visual scenes, resulting in new study topics [8]. There are two types of scene recognition approaches in the literature, depending on how a scene is classed for a picture. The top-down technique, which categorizes scenes based on picture attributes, is the first. The bottom-up technique, on the other hand, uses object recognition to determine the scene's category[9].

Feature-based scene recognition typically extracts discrimina- tive features from scene photos as the scene representation, which is then used to develop a semantic model based on that scene representation. Logistic Regression, K-means, Lin- ear Discriminant Analysis, and Support Vector Machine are examples of cutting-edge classifiers. The SVM classifier has been widely utilized for scene classification [10].

Visual object-based scene recognition can help separate very complex images that would otherwise be impossible to dis- tinguish using typical feature-based approaches. Superior per- formance on high-level visual identification tasks can be ac- complished with simple regularized logistic regression using Object Bank representation, in which an image is represented by integrating the image's response to several object detectors. Objects can be detected, and scenes may be classified using deep convolutional neural networks system topologies [11].

2.2 Deep learning based scene understanding

With the rapid progress of machine learning techniques, scene understanding has become more efficient and prefer- ment. In this section, we presents the basic architecture of deep learning (DL) based scene understanding. In addition, we investigate the most relevant DL-based models.

2.2.1 Basic architecture of DL-based scene understanding

Deep learning (DL) provides powerful techniques that dealwith complex tasks in image recognition and understanding.It offers computational models of multiple processing layersto extract relevant features and perform classification tasks. Several DL-based architectures has been proposed such as Convolutional Neural Networks (CNNs), Radial Basis Function Networks (RBFNs),Deep Belief Networks (DBNs),Multilayer Perceptrons (MLPs), Recurrent Neural Networks (RNNs), etc. However, convolutional Neural Networks (CNNs) are themost used DL- based architecture in image understanding field.CNNs are designed to recognize visual content directly frompixel of input images with minimal pre-processing. The basic architecture of a CNN is based on tree types of neural layersincluding (1) convolutional layers, (2) pooling layers, and (3)fully connected layers as shown in figure 1.

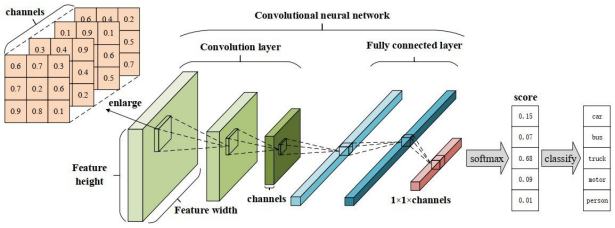

Fig. 1 Basic architecture of a basic convolutional neural network (CNN)[12]

The convolution task consist on the computation of feature maps basing on multiple kernels which slide through the wholeinput image. It enables faster learning training models.

The pooling layers allows minimizing the dimensions of feature maps by selecting the relevant features, basing on subsampling or downsampling to avoid loss of information by reducing size leads. There are two types of pooling: max and average pooling.

The fully connected layers performs high-level reasoning in neural network by applying several convolutional and pool- ing layer. Each neuron in a fully-connected layer is fully connected- to every other neuron in the previous layer. Thus, it used towards the end of a CNN in order to make predictions basing on the features learned by the previous layers.

In fact, the architecture of the CNN determines its perfor- mance and its efficiency basing on (1) the structure of layers,

(2) how elements are designed, and (3) which elements are present in each layer.

2.2.2 DL-based models used in scene understanding

Several DL-based models have been proposed and implemented in the last years. We presents in the following the most models applied on scene understanding.

LeNet: Proposed by LeCun et al. (1998) [13]. It is one of the most used CNN architectures. Five convolution layers and two fully connected layers compose the model. In order to reduce the spatial size of images, the max pooling parameter is applied between convolutional layers.

AlexNet: Proposed by Krizhevsky et al. (2012) [14], composed by five convolutional layers, with a combination of max- pooling layers, 3 fully connected layers, and 2 dropout layers. It applies a rectified linear unit (ReLU) which enables training models with simpler and faster computations.The activation function used in the output layer is Softmax. The total number of

parameters in this architecture is around 60 million.
ZFNet: Developed by Zeiler and Fergus (2013) [15], it is considered as refined architecture of the AlexNet. It applies a filter 7 X 7 o f s i z e . It is composed by seven layers.
VGGNet: Proposed by Simonyan and Zisserman (2015) [2].It uses small convolution filters (3 x 3) which increases the depth of the network and reduces the dimension of input dataat each laye. It can be with 16-layer (VGG-16) or 19 layers (VGG-19), with up to 95 million parameters and trained on over one billion images.
GoogleNet: Developed by Szegedy et al. (2015) [16], also named Inception V1, based on the LeNet architecture, includes twenty-two layers of smaller groups of convolutions known asinception modules, which uses the average pooling. The softmax function can predict the class score in the network.
ResNet: He et al. (2016) proposed a residual network called ResNet with 152 layers [1]. It is based on modules with 32 parallel paths. It proves that learning a residual function concerning layer input was more efficient than learning layer parameters without referring to inputs. It decreases the error function exponentially because of the use of skip connection in back-propagation. It uses eight GPUs for the training of the model.
DenseNet: Proposed by Huang et al. (2017) [3]. In this archi- tecture, all layers are directly connected with each other, which allows to reduce the number of parameters and strengthening feature propagation.
MobileNets: Developed by Howard et al. (2017) [17] for mobile device to classify images with low latency. They are considered as mall CNN architectures, which makes them easy to run in real-time using embedded devices like smartphones and drones. The structure of MobileNet is similar to VGG.
EfficientNet: Proposed by Tan and Le (2019) [18]. It based on scaling method of network depth, width, and resolution to obtain better performance.

2.2.3 related work

In the literature, several works have been proposed to deal with scene understanding using deep learning models. Authors in [19]proposed a strategy for reconstructing a 3D indoor scene without knowing the camera's internal calibration. The method used spatial rectangle projection to not only estimate the room layout of an image, but also to rebuild good details of the scene. The research measured the percentage of pixels categorized properly using SUN397 by comparing the room layout suggested to the room box ground truth. The system was capable of recreating various structures of indoor settings, and its accuracy according to the experiments.

The research presented in [20] introduced a novel object detector-based scene recognition algorithm using a histogram of items termed Bag of Objects. A simulation experiment was carried out with a large amount of image dataset named QVGA images acquired from the internet to test the proposed strategy. Consequently, it was discovered that the average scene recognition accuracy for 26 scene categories is 0.58.

A study by [21] proposed a hypothesis that terrain may be recognized effectively solely by measuring quantities related to the robot-ground interaction based on the NASA's Planetary Data System. Sensory signals are categorized as time series directly by a Recurrent Neural Network resulting from extra processing by a Convolutional Neural Network under these hypotheses. When compared to traditional Support Vector Machine, the results gained from genuine trials demonstrated equivalent or higher performance in both scenarios.

A research by [22] offered a low-cost method of indoor mobile robot navigation based solely on vision-based perception, re- ducing the challenge of visual navigation to scene classification using a convolutional neural network (CNN) with greater scene categorization accuracy and efficiency based on offline data collection and online navigation without additional sensors using only a camera. The system outperforms earlier relevant work in unknown contexts, according to both qualitative and quantitative results.

In another research by [23] the authors used common objects, such as doors or furniture, as a vital intermediate representation to recognize inside scenes as a distinguishing feature using collected images data set under a Creative Commons License from the image sharing SNS service. The proposed method is a generative probabilistic hierarchical model, in which low- level visual attributes are associated with objects via object category classifiers, and objects are associated with scenes via contextual linkages. The findings showed that the suggested method outperforms numerous state-of-the-art scene recogni- tion algorithms.

Moreover, a study by [24] presented a strategy that takes the scale into consideration and yields large recognition increases. Because the objects in the sceneries had their own range of scales, modifying the feature extractor to each scale was critical for improving recognition. Experiments on Delage's dataset demonstrated that recognition accuracy is significantly dependent on scale, and that using multi-scale combinations of ImageNet-CNNs and Places-CNNs could be increased to 66.26 percent.

3. Methodology and experimental results

3.1 Methodology overview

In order to build own scene understanding system that allows making robots to explore and discover indoor environ- ment, we suggest to use an existing DL-based model. However, since 1998, several models have been developed and tested on different areas, as described in section II-B2. Thus, we need to determine the most suitable DL-based model that deals with our context. To perform this, we suggest training significant DL-based models using existing image databases describing the indoor environment, including Places365 [4] , Scene15 [5], MIT Indoor67 [6], and ImageNet [7], as shown in figure 2.

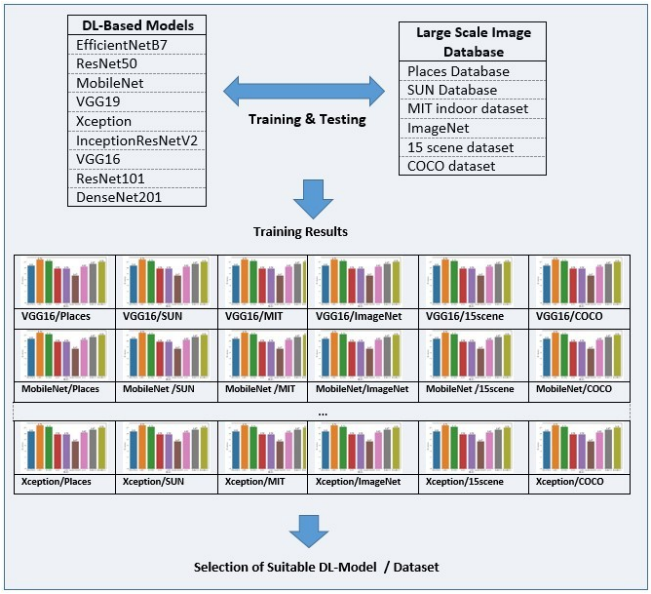

Fig. 2 Methodology overview

3.2 Experimental Results

We tested our methodology using Python programming language [25], keras library [26] and Google Colaboratory environment [27]. The figures 3, 4, 5, 6 show the results of training of DL-based models with Places365 [5], Scene15 [4], MIT Indoor67 [7], and ImageNet [6] respectively.

3.2 Discussion

For the Places365 image database, results show that the ResNet50 is the more appropriate model with accuracy of 76%. Besides, ResNet101 is the suitable model for the Scene15 image database with accuracy of 75%. Also, DenseNet201 gives best results with ImageNet Image database with accuracy of 92%. For the Indoor67 database, MobileNet is considered as the best model with accuracy of 80%. To summarize, Firstly, ResNet, DensNet and MobileNet DL-based models give better results than VGGNet, InceptionResNet, and EfficientNetB7. Secondly, DenseNet201 was the best DL-based model with the ImageNet Image database for indoor environment.

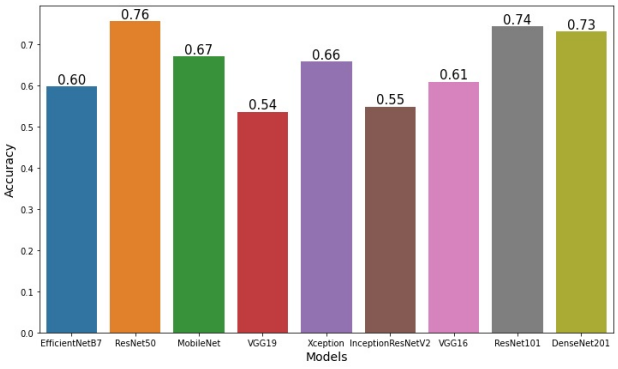

Fig. 3. Accuracy of training with Places365[4]

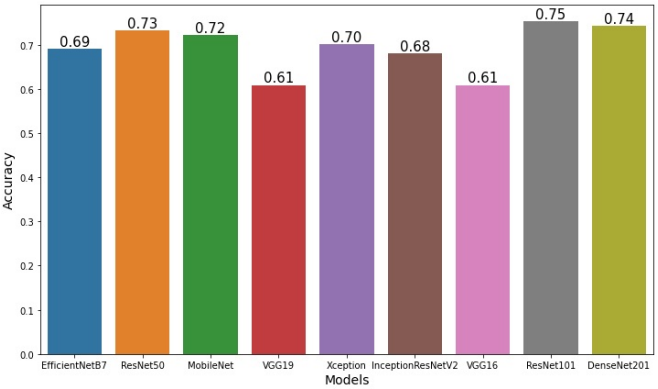

Fig. 4. Accuracy of training with Scene15[5]

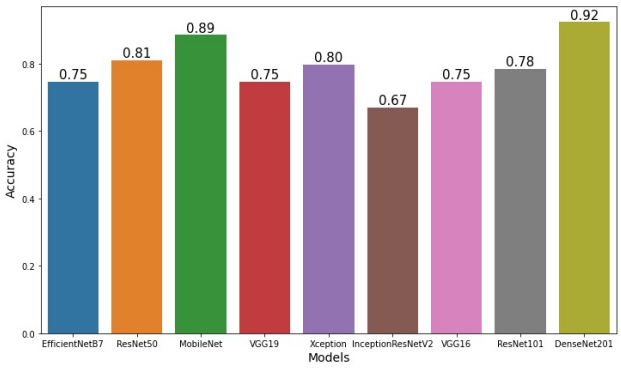

Fig. 5. Accuracy of training with ImageNet[7]

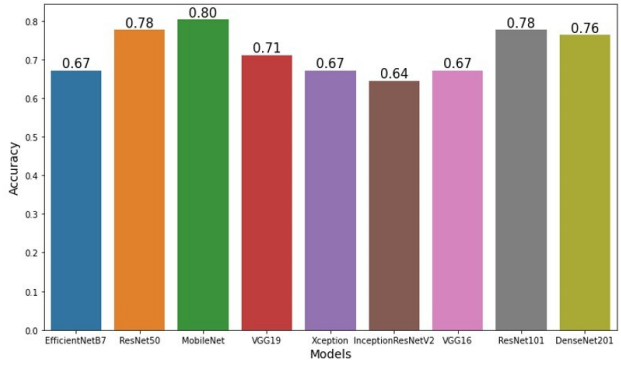

Fig. 6. Accuracy of training with Indoor67[6]

4. Conclusion

This paper aimed to present an overview of an intelligent scene recognition and understanding process based on several deep learning models and image databases. Thus, the chal- lenge was to identify the suitable image-database and deep learning model for building the scene recognition system of indoor environment. Therefore, we adopted a methodologythat starts to train significant DL-based models using existing image databases describing the indoor environment including Places365 [4] , Scene15 [5], MIT Indoor67 [6], and ImageNet [7]. Results showed that ResNet, DensNet and MobileNet DL-based models give better results than VGGNet, InceptionRes- Net, and EfficientNetB7. Besides, DenseNet201 was the best DL-based model with the ImageNet Image database for indoor environment.

References

[1] 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, 2016, pp. 770–778.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[3] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proceedings of the IEEE confer- ence on computer vision and pattern recognition, 2017, pp. 4700–4708.

[4] 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, 2017.

[5] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 2, 2006, pp. 2169–2178.

[6] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 413–420.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[8] C. Yan, L. Li, C. Zhang, B. Liu, Y. Zhang, and Q. Dai. (2019) Cross- modality bridging and knowledge transferring for image understanding.

[9] W. Yang, Y. Yuan, W. Ren, J. Liu, W. J. Scheirer, Z. Wang, T. Zhang,

[10] Q. Zhong, D. Xie, S. Pu et al., "Advancing image understanding in poor visibility environments: A collective benchmark study," IEEE Transactions on Image Processing, vol. 29, pp. 5737–5752, 2020.

[11] J. Wu, H. Zheng, B. Zhao, Y. Li, B. Yan, R. Liang, W. Wang, S. Zhou, G. Lin, Y. Fu et al., "Large-scale datasets for going deeper in image understanding," in 2019 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2019, pp. 1480–1485.

[12] C. Xu, Y. Dai, R. Lin, and S. Wang, "Stacked autoencoder based weak supervision for social image understanding," IEEE Access, vol. 7, pp. 21 777–21 786, 2019.

[13] X. Kang, B. Song, and F. Sun, "A deep similarity metric method based on incomplete data for traffic anomaly detection in iot," Applied Sciences, vol. 9, p. 135, 01 2019.

[14] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Handwritten digit recognition with a back-propagation network," Advances in neural information processing systems, vol. 2, 1989.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural informa- tion processing systems, vol. 25, 2012.

[16] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolu- tional networks. corr, abs/1311.2901,” arXiv preprint arXiv:1311.2901, 2013.
[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.
[18] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “Mobilenets: Efficient convo- lutional neural networks for mobile vision applications,” arXiv preprint arXiv:1704.04861, 2017.
[19] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for con- volutional neural networks,” in International conference on machine learning. PMLR, 2019, pp. 6105–6114.
[20] H. Wei and L. Wang, “Understanding of indoor scenes based on projection of spatial rectangles,” Pattern Recognition, vol. 81, pp. 497– 514, 2018.
[21] S. Masuda, Y. Kaeri, Y. Manabe, and S. Kenji, “Scene recognition method by bag of objects based on object detector,” in 2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS). IEEE, 2018, pp. 321–324.
[22] F. Vulpi, A. Milella, R. Marani, and G. Reina, “Recurrent and convo- lutional neural networks for deep terrain classification by autonomous robots,” Journal of Terramechanics, 2021.
[23] T. Ran, L. Yuan, and J. Zhang, “Scene perception based visual navi- gation of mobile robot in indoor environment,” ISA transactions, vol. 109, pp. 389–400, 2021.
[24] P. Espinace Ronda, A´ . M. Soto Arriaza, T. Kollar, and N. Roy, “Indoor scene recognition by a mobile robot through adaptive object detection,” 2013.
[25] 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, 2016, pp. 571–579.
[26] G. Van Rossum and F. L. Drake Jr, Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam, 1995.
[27] F. Chollet et al. (2015) Keras. [Online]. Available: https://github.com/fchollet/keras
[28] E. Bisong, Google Colaboratory. Berkeley, CA: Apress, 2019, pp. 59–64. [Online]. Available: https://doi.org/10.1007/978-1-4842-4470-87