Vision based Indoor Localization Method via Convolution Neural Network

Zeyad Farisi[1]
School of Automation Science and Engineering
South China University of Technology, Guangzhou, China
College of Community Service Department of Engineering and Science, Tabah University, Medinah, Saudi Arabia

Li Xiangyang[3]
School of Automation Science and Engineering
South China University of Technology
Guangzhou
China

Tian Lianfang[2]
School of Automation Science and Engineering, South China University of Technology; Research Institute of Modern Industrial Innovation, South China University of Technology; Key Laboratory of Autonomous Systems and Network Control of Ministry of Education
Guangzhou, China

Zhu Bin[4]
School of Mechanical and Electronic Engineering
Jiangxi College of Applied Technology
Ganzhou, China
School of Automation Science and Engineering
South China University of Technology
Guangzhou, China

***Abstract*—Existing indoor localization methods have bottleneck constraints such as multipath effect for Wi-Fi based methods, high cost for ultra-wide-band based methods and poor anti-interference for Bluetooth-based methods and so on. In order to avoid these problems, a vision-based indoor localization method is proposed. Firstly, the whole deployment environment is departed into several regions and each region is assigned to a location center. Then, in offline mode, the VGG16NET is pre-trained by ImageNet dataset and it is fine-tuned by images on a custom dataset towards indoor localization. In online mode, the fully trained and converged VGG16NET takes as input a video stream captured by the front RGB camera of a mobile robot and outputs features specific to the current location. The features are then used as input to an ArcFace classifier which outputs the current location of the mobile robot. Experimental results show that our method can estimate the location of a mobile object with imaging capability accurately in cluttered unstructured scenes without any other additional device. The localization accuracy can reach to 94.7%.**

***Keywords*—Indoor localization; VGG16NET; transfer learning; ArcFace classifier**

I. Introduction

With increasing demand of location based services, wireless location technology has gradually become a hot research topic. In outdoor environment, global positioning system (GPS), plough navigation systems and cellular positioning technology can satisfy the need of most of the localization. However, in the indoor environment, due to the block of building walls and items, the wireless signal transmission will be affected seriously which will prevent the accurate positioning badly. So other methods designed for indoor localization were proposed.

Among these methods, a strategy using Wi-Fi based fingerprinting location [1] is the most popular one. That is because Wi-Fi signal exists in almost every building in nowadays and it can be easily detected by cell phone which is already equipped almost for every person. So in the Wi-Fi based indoor localization system, additional device is not necessary. But this kind of methods have some critical defects: the value of Wi-Fi signal varies with time and easily affected by multipath effect. Although many algorithms have been proposed already for weakening the impact of these problems, none of them solve these problems well. Blue-tooth based methods [2] have good stability and are not be affected by multipath effect, but it is easily interfered by other blue-tooth devices. UWB (Ultra Wide Band) based methods [3] have strong anti-interference ability, high bandwidth and low power consumption, but its cost is high and the required device is expensive, what's more, the whole localization system is hard to deploy. The localization method based on RFID technology [4] and ZigBee technology [5] also need to deploy complex positioning system. So existing indoor location methods are either easily be disturbed or need support by complex auxiliary positioning system.

A vision based indoor localization method is proposed in the paper. The method uses images to determine the location of the moving object and has the following advantages: one, images in indoor scene are relatively stable than other wireless signals and are not interfered by other devices; two, the system only need a RGB camera on the detected object and no other complex device needed. The author in [6] proposed vision based indoor localization method using nature beacon and artificial beacon. In [7], author proposed a weighted KNN epipolar geometry-based approach for vision-based indoor localization. However, image matching based methods are susceptible to the shooting Angle, illumination changes and other unfixed image information. To find deep layer location features in image, [8] proposed a CNN (Convolutional Neural Network) based semantic scene segmentation for indoor robot navigation. Building information modelling (BIM) was used to construct the image dataset in [9] and a pre-trained

VGG16NET was applied to image feature extraction. But not all the buildings have BIM image dataset. What's more, [9] uses clean scene images for location and it is not suitable for practical application.

On the basis of [9], two major improvements were proposed in this paper. One, in the off-line mode, the VGG16 NET was pre-trained by ImageNet dataset first then fine-tuned by images labeled by location number, this Transfer learning technique allows for over-fitting to be addressed and achieve a robust model with only little training data. Two, original VGG16NET uses softmax as classifier, but when feature confounding occurs, the result will be wrong, for coping with this problem, ArcFace classifier [10] is applied to substitute Softmax classifier.

The rest of paper is organized as follows: Section 2 explains how the proposed algorithm works. Specifically, section 2-A describes the flowchart of our algorithm, section 2-B introduces how we use the VGG16NET, section 2-c explains why we use Arcface classifier. The experimental procedures and results are discussed in section 3. Conclusion and discussion in which show the contribution and limitation of our algorithm are presented in the end of the paper.

II. Vision Based Indoor Localization Method

A. The Flowchart of our Algorithm

Vision based indoor location algorithm in the paper applies VGG16NET to extract the deep layer location feature of images then the image feature was used to determine the location of the mobile object. Firstly, the whole scene is departed into several regions and each sub-region set a location number, the center of the sub-region is memorized. Then, photographs were taken with different angles around the location center. These photographs labeled by location number are then used as the training samples of our VGG16NET, our method include off-line mode and on-line mode.

The main purpose of the off-line mode is to construct the location feature in images of each location center. While training VGG16NET for feature extraction needs large number of images and our training samples are not enough. To avoid the over-fitting caused by the insufficient training dataset, transfer learning is employed. The VGG16NET is pre-trained by ImageNet dataset first then it is fine-tuned by our training samples. In on-line mode, image used for localization is put into the VGG16NET then we get the location feature of the input image. At last, ArcFace classifier was applied to classify locations for the input image. The flow chart of our algorithm can be seen in Fig. 1.

B. VGG16NET

The VGG16NET used in the paper consists of 13 convolution layers, 2 full connection layers and a classify layer. The 13 convolution layers were used to deepen the network then it can be used to extract the deep-layer image features, 5 max pooling layers are employed and they are used to reduce the feature dimensions in order to make the model computationally tractable. Transfer learning was used to improve the generalization ability of the model. The ImageNet dataset is applied to pre-train these first 13 convolution layers and weight parameters of in the network are adjusted and stored. The last three layers consist of two full-connection layers and a ArcFace layer. After the pre-trained step, the whole network is fine-tuned by images labeled with location number. In this step, all the weight parameters are adjusted. The structure of the CNN in the paper is shown in Fig. 2.

The convolution layers consist of 64 3*3 convolution kernels, batch normalization and Rectified Liner Uints (ReLU). The batch normalization was applied to limit large variances into a reasonable scope. In order to skip the unimportant information, max pooling layer which adopt a window with a stride size of 2 is used for down-sampling the feature maps. Full connected layers are employed to the combination of deep features with different weight parameters and random dropout layer are used to cut down some unimportant connection. At last, we can obtain the deep layer image location feature for each location number then we used ArcFace classifier to determine the location number for an input test image.

C. ArcFace Classifier

The original Softmax classifier in VGG16NET does not expand the edge of decision region which would cause wrong classification results when features of different class are similar. In vision based indoor localization, the location feature for adjacent location number may be similar, when the location feature is hard to distinguish, its positioning would be wrong. So the ArcFace layer is applied to substitute Softmax layer in this paper. The difference between the two classsier is mainly because of its loss function.

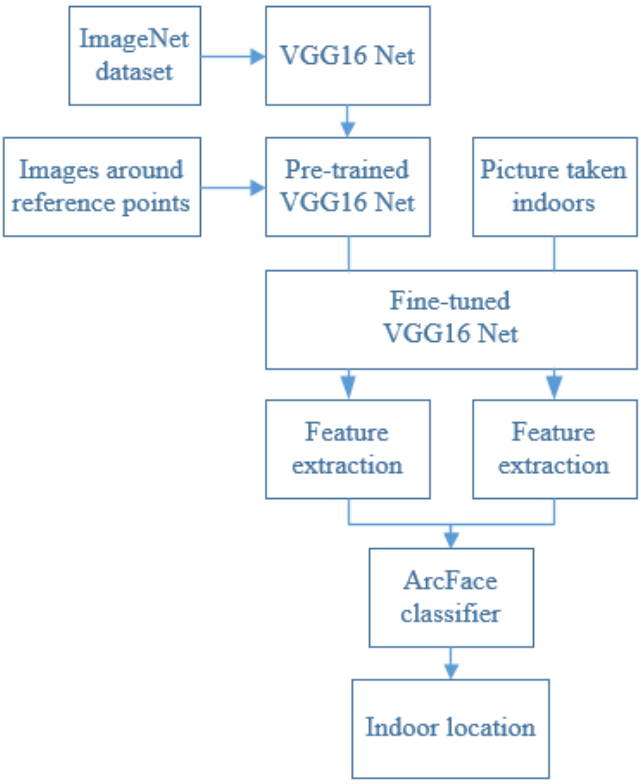


Fig. 1. Flowchart of Proposed Algorithm.

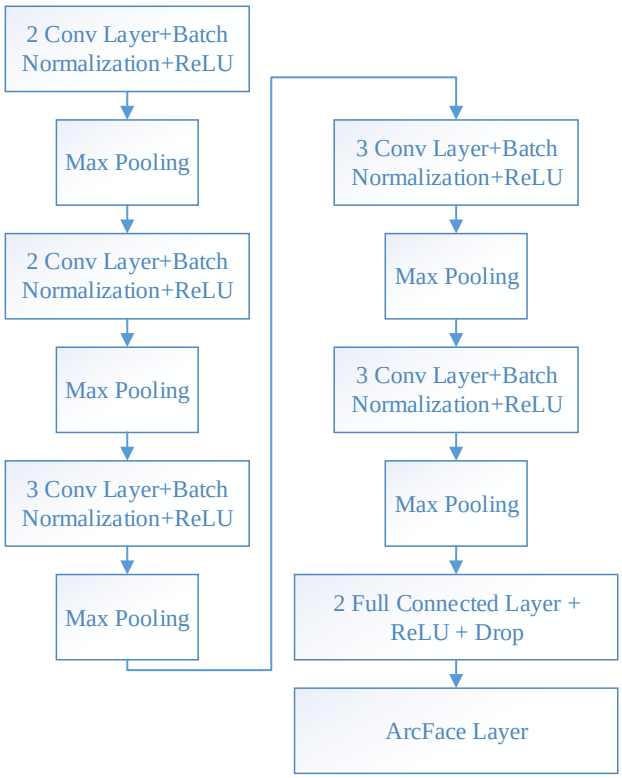


Fig. 2. Architecture of Proposed Convolutional Neural Network.

The loss function of Softmax is:

$$L_1 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} \tag{1}$$

Where x_i represents the feature of the i-th sample in the y_i-th class, W_j is the weight of the j-th column, b_j represents the bias term, N is the batch size and n is the class number. Based on (1), we let $b_j = 0$, $W_j^T x_i$ can transform to $\| W_j \| \| x_i \| \cos\theta_j$, θ_j is the angle between the weight W_j and the feature x_i, we let $\| W_j \| = 1$ and rescale $\| x_i \|$ to s by l_2 normalization. Then, we obtain the loss function as follows:

$$L_2 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos\theta_{y_i}}}{e^{s\cos\theta_{y_i}} + \sum_{j=1, j\neq y_i}^{n} e^{s\cos\theta_j}} \tag{2}$$

An additional angular margin penalty m was added into the angular between W_j and x_i to enhance intra-class compactness and inter-class discrepancy. We can obtain the loss function of ArcFace:

$$L_3 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos(\theta_{y_i}+m)}}{e^{s(\cos(\theta_{y_i}+m))} + \sum_{j=1, j\neq y_i}^{n} e^{s\cos\theta_j}} \tag{3}$$

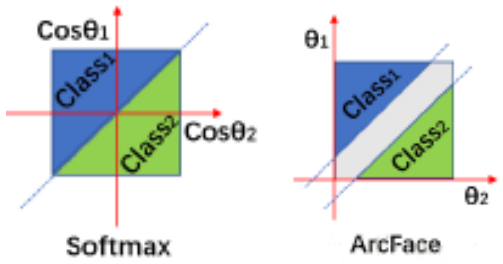


Fig. 3. Decision Margins of Two Loss Functions under Binary Classification Case.

Fig. 3 shows decision margins of two loss functions under binary classification case. The dashed blue line denotes the decision boundary. It can be seen that the Softmax loss function roughly separate the whole plan into two parts and no margin exists between decision boundary, while the ArcFace loss function can obviously afford decision margin between decision boundaries.

III. Experimental Result Analysis

Experiments are carried out in an industrial environment where the whole scene is departed into 9 regions and each region is assigned to a location center, respectively. The assignment of our vision based indoor localization algorithm is to determine the real-time location of the mobile robot in real time. Fig. 4 shows the floor plan of the experimental scene with 9 location centers. The mobile robot used in our experiment is consist of a RGB camera, motion control system and a laptop which is equipped with a single Nvidia GTX 1080 card and an Intel i7 processor. The appearance of the robot can be seen in Fig. 5.

Fig. 4. Floor Plan of the Experimental Scene.

Fig. 5. Experimental Robot Platform.

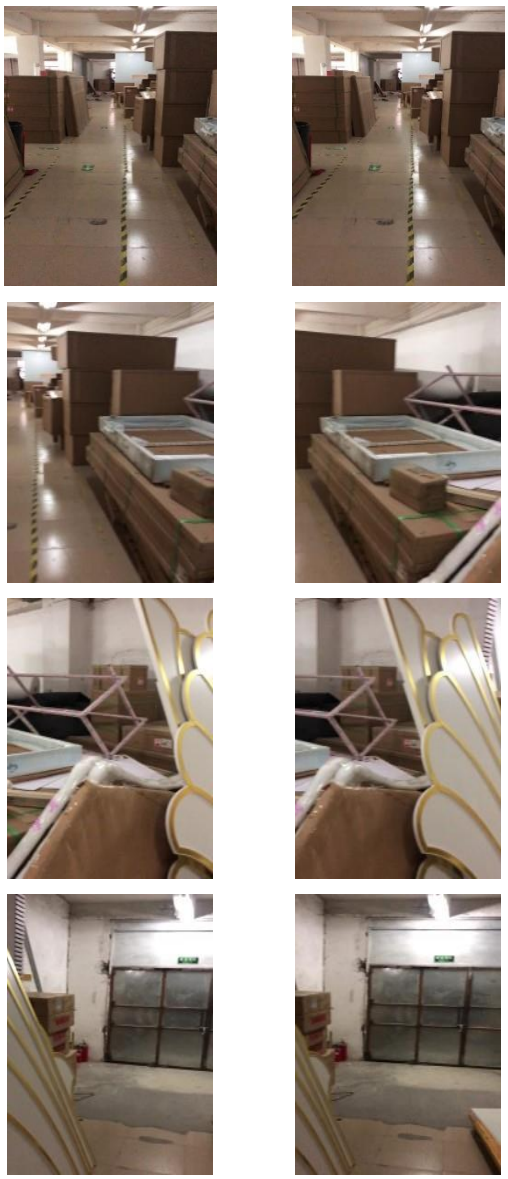


Fig. 6. Some of Training Samples of Location Center "1".

In order to cut down overall training time, the pre-trained VGG16NET is applied directly which can be downloaded in GitHub. The fine-tuning of the network is performed using images labeled location number, these images are taken around the location centers. 100 pictures of each location center were used for training. Fig. 6 shows some of the training samples in location "1".

Fig. 7 shows the change of the model accuracy during iteration processes. We can see that the model accuracy increases for both training set and test set when the iteration time increases. When the iteration time reach to 35, the accuracy of the model would change little during the iteration time increases. The highest accuracy of training set and test set reach to 95.2% and 94.3, respectively. Fig. 8 shows curves of the loss function for training set and test set. We can see that the value of the loss decreases for both training set and test set when the iteration time increases. The lowest loss of training set and test set reach to 0.03 and 0.05, respectively.

After finished the training of the VGG16NET, the model can be used for image based location classification. 21 images of each location center, totally 189 images were used as the experimental objects. Confusion matrix was used to depict the experimental results which were shown in Fig. 9. Fig. 9(a) is the classification result of ArcFace classifier, Fig. 9(b) is the classification result of Softmax classifier, rows in the figure mean the actual location of tested images, and columns mean the estimated location tested by algorithms. When the estimated location meet the actual location, i.e. the check of left diagonal, means that the estimated location is right. Numbers in these checks mean the time of correct classification, numbers in other checks mean the time of wrong localization and on which location. For example, in the fourth line of Fig. 9(a), the actual location is location "3", the correct estimated time is 18, and the wrong estimated time is 3, one time in location "6" and two times in location "8". Totally, the correct classification time for Softmax classifier is 172 and the wrong classification time is 17, the accuracy rate of softmax classifier is 91%. When comes to ArcFace classifier, the correct classification time is 179 and the wrong classification time is 10. The wrong classification mainly concentrate on the center of the building where images taken in these location points are similar to images taken from other location points, sometimes even human eyes can't identify. The accuracy rate of ArcFace classifier reaches to 94.7% which is 3.7% higher than the Softmax classifier.

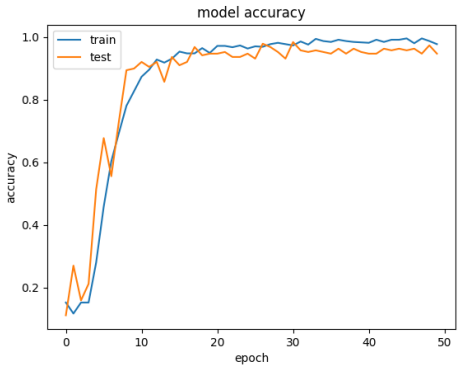


Fig. 7. Curvers of the Model Accuracy for Training Set and Test Set.

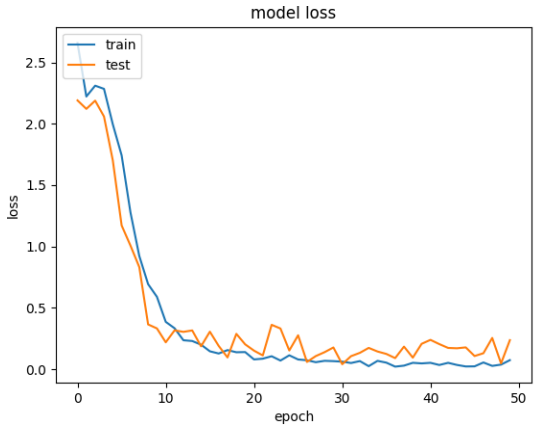
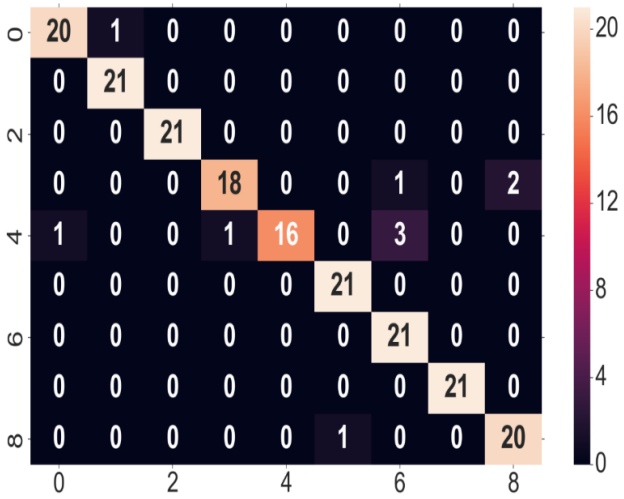
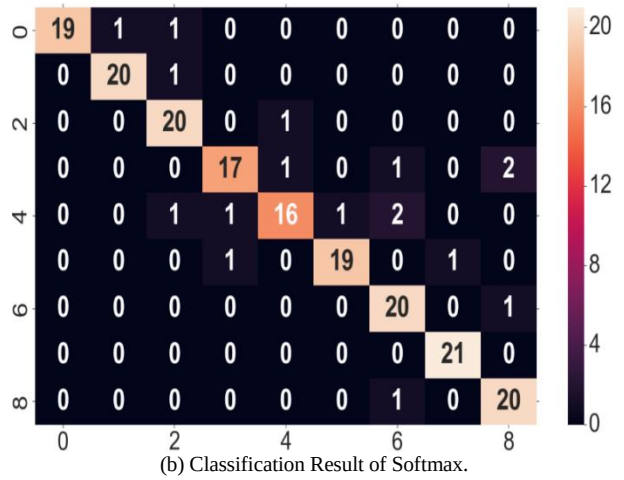


Fig. 8. Curvers of the Loss Function for Training Set and Test Set.



(a) Classification Result of ArcFace.



(b) Classification Result of Softmax.

Fig. 9. Expermental Result Depicted by Confusion Matrix.

IV. Conclusion and Discussion

A vision based indoor location method is proposed in this paper to avoid problems (easily be disturbed by other signals or complex devices needed) hard to be solved in existing indoor location systems. CNN is applied to extract the deep positioning feature in image to solve the problem of existing vision based method easily be disturbed by superficial image feature. Transfer learning was employed to pre-trained VGG16NET then it is finetuned by images labeled by location number. After that, ArcFace classifier is used to substitute Softmax classifier to solve the problem of the error classification in feature similar circumstances. Experimental results show that our method can acquire high accuracy location of object with imaging capability in cluttered unstructured scene. But when the location information of the input image is obscured badly, our method would fail, such as when a human stand closely in front of the camera. What's more, the positioning accuracy of the algorithm in this paper is not high, that is because if location centers are placed closely, location features of nearby location center will hard to identify.

References

[1] K. Kim, S. Lee, K. Huang. "A scalable deep neural network architecture for multi-building and multi-floor indoor localization based on Wi-Fi fingerprinting," Big Data Analytics, 1th ed, vol. 3, pp. 1-9, 2018.

[2] E. Cabrera, D. Camacho. "Towards a Bluetooth Indoor Positioning System with Android Consumer Devices," The IEEE International Conference on Information Systems and Computer science, Quito, Ecuador, pp. 56-59, 2017.

[3] P. Gong, Y. Mao, Y. Du. "An UWB indoor location algorithm based on wavelet de-noising," Microcomputer&Its Applications, 6th ed, vol.32, pp. 66-69, 2015.

[4] L. Qiu, Z. Huang, N. Wirstrom, T. Voigt. "3DinSAR: Object 3D localization for indoor RFID applications," The IEEE International Conference on RFID, Orlando, USA, pp. 1-8, 2016.

[5] Y. Wang, Y. Zhao, M. Li. "Application research of Gauss filter in Zigbee indoor positioning," Journal of Geomatics, 1st ed, vol. 5, pp. 512-517, 2016.

[6] W. Hong, H. Xia, X. An, X. Liu. "Natural Landmarks based localization algorithm for binocular vision," The 29th Chinese Control And Decision Conference, Chongqing, China, pp. 3313-3318, 2017.

[7] H. Sadeghi, S. Valaee, S. Shirani. "A weighted KNN epipolar geometry-based approach for vision-based indoor localization using smartphone cameras," The IEEE 8th Sensor Array and Multichannel Signal Processing Workshop, Coruna, Spain, pp. 37-40, 2014.

[8] D. Bersan, R. Martins, M. Campos, et al. "Semantic map augmentation for robot navigation: A Learning Approach Based on Visual and Depth Data, " The IEEE Latin American Robotic Symposium, Joao Pessoa, Brazil, pp: 45-50, 2018.

[9] F. Walch, C. Hazirbas, T. Sattler, et al. "Image-based localization using LSTMs for Structured Feature correlation," The IEEE International Conference on Computer Vision, Venice, Italy, pp. 627-637, 2017.

[10] J. Deng, J. Guo, S. Zafeiriou. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," Computer Vision and Pattern Recognition, 6th ed, vol. 32, pp. 1-11, 2018.