The Image-Based Positioning Technique Using Inter-Pixel Relation Network

ZHENYU ZHANG1, 2, (Member, IEEE), AND QIULING ZHAO1, (Member, IEEE)

1Liaoning Institute of Science and Technology, Benxi 117004, China
2School of Electrical and Information Engineering, Liaoning Technical University, Fuxin 125105, China

Corresponding author: Zhenyu Zhang (442913397@qq.com)

This work was supported in part by the Natural Science Foundation of China under Grant NSFC21702086, in part by the Natural Science Foundation of Liaoning Province of China under Grant 20180551039, and in part by the Doctoral Scientific Research Foundation of Liaoning Institute of Science and Technology under Grant 1910B10.

ABSTRACT Accurate location information is very important for human beings and agents. However, GPS has very limited positioning accuracy or even can not work in challenging environments such as indoor and urban canyons, and other solutions suffer from different shortcomings. It is very meaningful to use the scene image captured by the high-definition camera on the smart phone to locate indoor through the neural network, but the existing image-based positioning methods do not take into account the interrelationship of different instances of the same feature space. In this paper we present a novel network architecture, which embeds Scene Enhancement module and Relational Network (RN) into the end-to-end positioning system. The system was tested in three typical scenarios, and the test results show that its performance outerperforms existing image-based schemes in most cases.

INDEX TERMS Indoor positioning, inter-pixel relation network, scene image enhancement.

I. INTRODUCTION

With the development of science and technology, accurate location information becomes more and more important for human beings or other agents. Global positioning system (GPS) provides an effective means of outdoor positioning, but there are obstacles to the line of sight signal between the satellite and the receiver, and the signal are attenuated by buildings and outdoor obstacles, GPS shows very limited accuracy or even can not locate in urban canyon, close to wells, indoor and underground environments [1].

Several indoor positioning solutions have been proposed. One solution is Infrared radiation (IR) which use wireless technology, but has security and privacy issues, as well as interference from fluorescent lights and sunlight, and high hardware and maintenance costs [1]. Bluetooth is considered another promising indoor positioning technology, the downside of this method is its increased latency and power consumption, which makes it unsuitable for real-time positioning [2], [3]. Ultrasonic can also be used for indoor positioning, but it cannot penetrate walls, has low accuracy and is vulnerable to metal collisions and other sources of interference [4], [5]. WLAN is the most widely used indoor positioning means, but phase sampling requires a priori knowledge of existing access points, which is very time consuming. The coordinates, orientation, Gaussian distribution and corresponding wireless access points of the collected signals add a great deal of manual labor, and the accuracy is reduced due to the differences in the signals [6]. SLAM can construct three-dimensional maps of unknown indoor environments and then infer the location of the agent in the map. However, the amount of computation required to build a 3D model is very large, requiring laser rangefinders, RGB-D sensors and cameras to do the computation simultaneously, so it is not practical for indoor navigation [7].

The above solutions can only be achieved by installing expensive electronic devices. Smartphones are now equipped with high-definition cameras, and image-based positioning technology can significantly reduce deployment costs, which can be achieved through convolutional neural networks (CNN). CNN is a feedforward neural network in which artificial neurons respond to surrounding neurons, which has excellent performance for large format image processing. It is of great interest for predicting camera position and orientation using convolutional neural networks in urban canyons,
indoors and other environments where satellite navigation is inaccurate or even fails.

PlaNet [8] treats image localization as dividing the earth’s surface into thousands of multi-scale geographic units, and trains a deep network using millions of geo-tagged images. Unlike previous methods of identifying landmarks using global image descriptors, this model integrates multiple visible cues. This scheme has achieved better results than previous methods, and even reaches superhuman accuracy in some cases. In addition, the model is combined with Long Short-Term Memory by learning to use temporal coherence to locate indeterminate photographs. Its performance is improved by 50% over a single image model. Posenet [9] propose a robust real-time monocular six-degree-of-freedom relocation system, by training a 23-layer convolutional neural network. The system returns the 6-DOF camera pose from a single RGB image in an end-to-end manner without the need for additional engineering or graphics optimization. The system can run in real time at a speed of 5 milliseconds per frame, with an accuracy of about 2 meters 6 degrees for large outdoor scenes and about 0.5 meters 10 degrees for indoor scenes. In [10], a new CNN+LSTM architecture is proposed for camera pose regression in indoor and outdoor scenes. The CNN has robustness to motion blur and illumination changes, and the use of LSTM units at the output of the CNN to reduce the structured dimensions of the feature vectors leads to a significant improvement in positioning performance. In addition, a new large indoor data set with ground truth is presented, and indoor and outdoor experimental results show that this method is superior to previous methods and can be effectively positioned on mostly untextured surfaces.

Although CNN has made great progress in handling unstructured data, it does not consider the relationship between different instances in the same scene image, which can be solved well using RN, and has achieved good results in visual QA and test-based QA [11]. RN has a structure suitable for inference and can infer the logical relationship between any two instances. In addition, RN is more efficient in that it uses one function to compute all relationships and therefore has a strong generalization capability.

Instance segmentation enables object detection and pixel-level segmentation, which is a combination of object detection and semantic segmentation, with better performance than object detection and semantic segmentation, but more difficult to implement than both. Instance segmentation can detect the edge of an object, whereas object detection can only be accurate to the boundary box; instance segmentation can distinguish between different instances of the same class in a graph, whereas semantic segmentation cannot. Jiwoon Ahn proposed a novel neural network architecture that is trained to provide two features, namely instance map and pairwise affinities [12].

Specifically, we propose an end-to-end positioning system that utilizes the two features and RN described above. First, a Scene Image Enhancement module was developed. Secondly, instance map and pairwise affinities are generated using the IRNet module in [12]. Next, the parameters of the above two trained modules are uploaded to the corresponding one of the end-to-end positioning system. Finally, the performance of the entire system is evaluated.

II. BACKGROUND

A. CNN

CNN is a deep feedforward neural network with convolution computing unit, and has been applied to computer vision, natural language processing and other fields. With CNN become deeper, vanishing gradient should be considered. With the gradient propagates back to the front layer, repeated multiplication may make the gradient infinitesimal, resulting in saturation or even a decrease in accuracy. Residual Neural Network (ResNet) solves this deficiency well [13], ResNet has a deeper network architecture than VGGNet, but fewer parameters. When several unique modules called identity shortcut connection are embedded, the deeper network should not lead to the increase of error on the training set.

B. RELATION NETWORK

Reasoning is an essential ability for artificial intelligence, but it’s not realistic for CNN. The RN module can be inserted into the CNN to enable it to handle unstructured inputs and to infer logical relationships between instances. During the training process, the RN-enhanced network first understands the entire image and identifies the instances into different categories. Then use the RN module to establish relationships and add all these relationships to generate output for all instance pairs in the image.

III. MODELS FOR IMAGE BASED POSITIONING SYSTEM

As shown in Fig. 1, the proposed end-to-end positioning system consists of Scene Image Enhancement, Inter-pixel Relation Network, Pixel-level Features, G-MLP and F-MLP. The entire system is trained in end-to-end manner, i.e., there is no need to manually create labels for the training image, and regression is performed directly on the image itself. The idea behind it is to make the system get the position and orientation of camera from the “most primitive” image. The end-to-end learning reduces human workload and performs better in most applications. In the following sections, we will describe in detail the composition of each module and how it works.

A. SCENE IMAGE ENHANCEMENT

Image preprocessing is necessary to eliminate irrelevant information, recover useful real information and enhance the

FIGURE 1. The pipeline of proposed positioning system.
data.

\[ D_{\text{rev}}(\tilde{I}) = \text{sigmoid}(C(I_g) - C(I_r)) \]  

(2)

The loss functions of Discriminator and Generator can be written as:

\[ L^G_{\text{SGAN}} = -E_{(I_r, I_g) \sim (P, Q)}[\log(\text{sigmoid}[C(I_r) - C(I_g))])] \]

\[ L^D_{\text{SGAN}} = -E_{(I_r, I_g) \sim (P, Q)}[\log(\text{sigmoid}[C(I_g) - C(I_r))])]. \]

(3)

B. INTER-PIXEL RELATION NETWORK

The Inter-pixel Relational Network (IRNet) is similar to the one in [12], except that RReLU is used as the linear rectifier unit. IRNet has two distinct branches, Instance Map and Pairwise Affinities, which are pixel-level features that distinguish between different objects belonging to the same class and can be used as features of end-to-end positioning systems. The two branches share the same ResNet backbone, each branch can extract features from five levels of the backbone, and generate feature through their own branches. The convolutional layers of the two branches are followed by Batch Normalization and RReLU, except for the last layer, which is detailed in [12]. And the weights of IRNet can be transferred from the trained IRNet to the end-to-end positioning system.

C. G-MLP AND F-MLP

Instance map and pairwise affinities are concatenated and fed into Multilayer perceptron (MLPs), respectively. The following composite functions can be used to calculate object relationships [16]:

\[ R(O) = f_\theta(\sum_{i,j} g_\theta(o_i, o_j)) \]  

(4)

where \( o_i, o_j \) are the \( i \)th and \( j \)th object in the concatenated features. \( f_\theta \) and \( g_\theta \) are MLPs, its parameters are learnable, and the output of \( f_\theta \) indicate the ways in which two objects are related.

IV. EXPERIMENTS AND RESULTS

We train and evaluate the performance of the proposed framework on three typical datasets, Cambridge outdoor scenes [9], small-scale indoor scenes [17] and new TUM-LSI dataset [10]. The first kind of dataset covers a wider range of space, and is a perfect scene for image based localization. The second method has a larger number of images in a smaller spatial range and is more suitable for augmented reality. The spatial coverage of the third dataset is twice as large as that of the second dataset and allows for a more efficient assessment of the robustness of indoor positioning performance. Two thirds of the three augmented dataset are randomly selected as the training set and the rest as the test set.

A. EXPERIMENTAL SETUP

The Generator, Discriminator and RN we used in this case is detailed in Table 1. In order to reduce the number of
parameters and the number of calculations, the remaining blocks were designed using a bottleneck approach. For the middle $3 \times 3$ convolutional layer, it first uses the previous $1 \times 1$ one to reduce the amount of computation and then uses the subsequent $1 \times 1$ one to restore. The depth of all three architectures is 45 and each contains five modules: conv1 to conv5. Each module in the Discriminator performs convolution and Maximum-pooling and stores the corresponding Maximum-pool indices. Instead, the Generator upsamples using the corresponding Max-pooling indices in the Discriminator and then performs convolution.

**B. TRAINING DETAILS**

The proposed system is trained using the same loss function used in [9] directly to estimate the camera pose from the scene image, and the pose vector $p$ is represented by the three-dimensional camera position $x$ and the direction represented by the quaternion $q$.

$$\text{loss}(I) = ||\hat{x} - x||_2 + \beta||\hat{q} - q||_2$$

All experiments are performed on Dell server equipped with NVIDIA P4000 GPU using Pytorch with ADAM optimizer. The training process generally consists of three steps: GAN initialization, IRN initialization and System-wide training.

Firstly, the input image randomly selected image from training dataset is flipped, rotated and fed to Discriminator as mentioned earlier. In the process of training Generator and Discriminator, it usually necessary to fix the weights of one branch, update the weights of another branch, and then iterate alternately. Both branches try to optimize their own network, resulting in a competitive confrontation until both sides reach a dynamic balance. The generated image with the corresponding input image’s positional information constitutes the complete training data and is added to the existing training data set.

Secondly, the proposed relation network in table 1 is used to replace the one in instance segmentation system [12]. And the replaced relation network is trained using the same method in [12].

Finally, the weights of trained Generator, Discriminator and RN are transferred to the one in the proposed end-to-end positioning system. A randomly selected image from the enhanced training dataset is fed into the end-to-end positioning system, and an error is obtained by subtracting the predicted position and orientation from the ground truth. The error propagates backwards in the model, and each layer adjusts the weights according to the error. If the system does not converge, another image is selected for input to the positioning system, and the above regression process is repeated until the model converges or reaches the desired effect.

**C. EXPERIMENTAL RESULTS**

The CNN based [9] and CNN+LSTM based location method [10] are compared with the proposed one in this paper. The hyperparameter settings are as follows: crop size 512, batch size 32, learning rate 0.1 and weight decay $1e^{-1}$.

In Table 2, we present the large scale outdoor positioning results of the three methods which are trained and evaluated using enhanced Cambridge Lanadmark dataset, which includes four typical scenarios such as King’s College, Old Hospital, Shop Facade and St Mary’s Church. The average position accuracy of the proposed method is 80.9% and 5.4% higher than that of the former two methods, and the average orientation accuracy is increased by 20.8% and 1.9% respectively. The small scale indoor positioning results evaluated on 7 Scenes dataset are presented in Table 3, we can conclude that our methods outperform posenet and CNN+LSTM. The average error of position is reduced from 0.44m, 0.31m to 0.28m, while the average error of orientation is reduced from 10.4°, 9.85° to 9.55°. Finally, evaluation results on TUM-LSI in Table 4 show that our method is comparable to former two methods.
V. CONCLUSION

We propose a novel image-based end-to-end positioning system, which draws on the latest results of instance segmentation, generates the relationship between different instances in the same feature layer, and embeds the scene Image Enhancement module and the IRNet module into the end-to-end positioning system. The test results in three typical scenarios show that our method is better than or comparable to PoseNet and CNN + LSTM.

In addition to indoor positioning, the proposed method can also be used for autopilot, and has a high cost-performance ratio compared to the existing one. However, due to the complex neural network architecture, its implementation in hardware will be more difficult, requiring further network pruning. Furthermore, the positioning system based on “CNN + LSTM + RN” may achieve better performance than the proposed one, but due to the limited space, it should be further studied in the subsequent work.

REFERENCES

[1] Y. Gu, A. Lo, and I. Niemegeers, “A survey of indoor positioning systems for wireless personal networks,” IEEE Commun. Surveys Tuts., vol. 11, no. 1, pp. 13–32, 1st Quart., 2009.

[2] F. Subhan, H. Hasbullah, A. Rozzyev, and S. T. Bakhsh, “Indoor positioning in Bluetooth networks using fingerprinting and lateration approach,” in Proc. Int. Conf. Inf. Sci. Appl., Apr. 2011, pp. 1–9.

[3] H. J. P. Iglesias, V. Barral, and C. J. Escudero, “Indoor person localization system through RSSI Bluetooth fingerprinting,” in Proc. 19th Int. Conf. Syst., Signals Image Process. (IWSSIP), Vienna, Austria, Apr. 2012, pp. 40–43.

[4] A. Runge, M. Baunach, and R. Kolla, “Precise self-calibration of ultrasound based indoor localization systems,” in Proc. Int. Conf. Indoor Positioning Indoor Navigation, Sep. 2011, pp. 1–8.

[5] C. Medina, J. C. Segura, and S. Holm, “Feasibility of ultrasound positioning based on signal strength,” in Proc. Int. Conf. Indoor Positioning Indoor Navigation (IPIN), Nov. 2012, pp. 1–9.

[6] S. A. Ahson and M. Ilyas, Location-Based Services Handbook: Applications, Technologies, and Security, 1st ed. Boca Raton, FL, USA: CRC Press, 2010, pp. 67–90.

[7] P. Chechcin, F. Géronnier, C. Blanc, R. Chapuis, and L. Trassoudaine, “Location-based services handbook: Applications, technologies, and security,” in Tracts in Advanced Robotics, vol. 62. Boca Raton, FL, USA: CRC Press, 2010, pp. 151–161.

[8] T. Weyand, I. Kostrikov, and J. Philbin, “PlaNet—Photo Geolocation with convolutional neural networks,” in Proc. Eur. Conf. Comput. Vis. New York, NY, USA: Springer, 2016, pp. 37–55.

[9] A. Kendall, M. Grimes, and R. Cipolla, “PoseNet: A convolutional network for real-time 6-DOF camera relocalization,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 2938–2946.

[10] F. Walch, C. Hazirbas, L. Leal-Taixe, T. Sattler, S. Hilsenbeck, and D. Cremers, “Image-based localization using LSTMs for structured feature correlation,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 627–637, doi: 10.1109/ICCV.2017.75.

[11] A. Santoro, D. Raposo, and D. G. Barrett, “A simple neural network module for relational reasoning,” in Proc. Adv. Neural Inf. Process. Syst. NIPS, 2017, pp. 4967–4976.

[12] J. Ahn, S. Cho, and S. Kwak, “Weakly supervised learning of instance segmentation with inter-pixel relations,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 2209–2218.

[13] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” 2015, arXiv:1511.06434. [Online]. Available: http://arxiv.org/abs/1511.06434

[14] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[15] B. Xu, N. Wang, T. Chen, and M. Li, “Empirical evaluation of rectified activations in convolutional network,” 2015, arXiv:1505.00853. [Online]. Available: http://arxiv.org/abs/1505.00853

[16] A. Santoro, D. Raposo, D. G. T. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, “A simple neural network module for relational reasoning,” 2017, arXiv:1706.01427. [Online]. Available: http://arxiv.org/abs/1706.01427

[17] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon, “Scene coordinate regression forests for camera relocalization in RGB-D images,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2013, pp. 2930–2937.

ZHENYU ZHANG (Member, IEEE) received the M.Sc. and Ph.D. degrees in electronic engineering from the University of Chinese Academy of Sciences, China, in 2004 and 2010, respectively. In 2004, he joined the Electronic Engineering Laboratory and the China Academy of Engineering Physics, Mianyang, Sichuan, as a Research Associate. Since 2019, he has been an Associate Professor of telecommunication engineering with the Liaoning Institute of Science and Technology, Benxi. His current research interests include modulation classification and parameter estimation, artificial intelligence, and VLSI design. He is also a Project Collaborator of the National High Technology Research and Development Program of China.

QIULING ZHAO (Member, IEEE) received the Ph.D. degree in bioinformatics from the University of Chinese Academy of Sciences, China, in 2014. Since 2019, she has been an Assistant Professor of bioinformatics with the Liaoning Institute of Science and Technology. Her current research interest includes signal processing in gene synthesis. She has been a Principle Investigator on several National Nature Science Foundation of China projects.

* * *