KrNet: A Kinetic Real-Time Convolutional Neural Network for Navigational Assistance

Shufei Lin [1][0000-0003-4911-9443], Kaiwei Wang[1], Kailun Yang[1], and Ruiqi Cheng[1]

[1] State Key Laboratory of Modern Optical Instrumentation, Zhejiang University, Hangzhou, China
wangkaiwei@zju.edu.cn

Abstract. Over the past years, convolutional neural networks (CNN) have not only demonstrated impressive capabilities in computer vision but also created new possibilities of providing navigational assistance for people with visually impairment. In addition to obstacle avoidance and mobile localization, it is helpful for visually impaired people to perceive kinetic information of the surrounding. Road barrier, as a specific obstacle as well as a sign of entrance or exit, is an underlying hazard ubiquitously in daily environments. To address the road barrier recognition, this paper proposes a novel convolutional neural network named KrNet, which is able to execute scene classification on mobile devices in real time. The architecture of KrNet not only features depthwise separable convolution and channel shuffle operation to reduce computational cost and latency, but also takes advantage of Inception modules to maintain accuracy. Experimental results are presented to demonstrate qualified performance for the meaningful and useful applications of navigational assistance within residential and working area.

Keywords: Convolutional Neural Network, Scene Classification, Mobile Navigational Assistance, Visually Impaired.

1 Introduction

According to the World Health Organization, 253 million people are estimated to be visually impaired and 36 million people are totally blind all over the world [1]. The most critical navigation task for people with visually impairment is to reach a destination without colliding with obstacles. Towards this end, mobile localization plays an important role beyond obstacle avoidance. In some general mobile navigational applications, the outdoor positioning error is between 3 to 10 meters, and it is even worse under some severe weather conditions. Visual place recognition provides valuable information to enhance situational awareness. It is important noting that the road barrier, which is designed to limit the passage of vehicles on the road, is usually set at the gate of a residential area or working area, as shown in Figure 1. The road barrier could be taken as a sign of the entrance or exit. For visually impaired people, the difficulty in these scenarios is that the road barrier can be bypassed instead of being avoided, which is different from ordinary obstacles. Thereby, barrier recognition is clearly desirable to

complement general assistance systems featuring obstacle avoidance or mobile localization.

As is known to all, convolutional neural networks have not only achieved remarkable capabilities in both computer vision [2-4] and robotics communities [4] through years of research, but also been applied to visual place recognition [5] to enhance situational awareness. Following this line, we focus on the recognition of road barrier and dedicate to providing assistance based on CNN within specific scenarios where visually impaired people travel in daily environments, such as residential area or working area.

To address the road barrier recognition, this paper proposes a light weight and efficient convolutional neural network named KrNet. The architecture of our network, based on depthwise convolution and channel shuffle operation, has been designed to maximize its performance and keep efficiency that is suitable for real-time inference on a portable CPU. We evaluate the proposed network in navigational assistance within residential community, describing the complete applied process of our multi-sensor system to assist visually impaired people in real-world scenarios.

(a) (b) (c) (d)

Fig. 1. Classification results in real-world scenarios. The images are classified as two classes, namely "barrier" and "other" (non-barrier), with the corresponding classification confidence value. (a) Road barrier set at the entrance of a working area. (b) Road barrier set at the entrance of an underground parking lot. (c) (d) Scenarios without road barrier are classified correctly as background even though the texture of curbs is similar to the texture of road barriers.

2 State of the Art

Recent researches on deep convolutional neural networks have concentrated on improving the classification accuracy. Benefit from large datasets (e.g. ImageNet [6]), powerful hardware and improved algorithm, AlexNet [2] eventually achieved its success in 2012. To further promote the accuracy, a straightforward way is to increase the depth and width of networks. VGG [7] replaces a 7×7 filter with a few 3×3 filters but still simply stacks standard convolutional layers. GoogLeNet [8] proposes Inception module which simultaneously uses four filters with different kernel size to extract feature. ResNet [9] utilizes the bypass connection solving the notorious problem of vanishing gradients to achieve impressive performance. However, all of the networks above have large depth and size, which poses a huge challenge for deploying these deep learning algorithms on the mobile devices with limited computational resources. The compression and acceleration of CNN models have become one of the most important research fields in both academia and industry. Deep compression [10] makes use of pruning, vector quantization and Huffman encoding to compress weights. Distilling

[11] uses a larger pre-trained network and then transfers it to train a smaller network. SqueezeNet [12] is based on the Fire module which is comprised of a squeeze convolution layer and an expand layer. MobileNet [13] uses depthwise separable convolutions to reduce computational cost. A channel shuffle operation is come up with to allow input and output channels to be related with each other [14]. To the best of our knowledge, all these networks have not been used to aid visually impaired individual in navigation.

For the visually impaired, we have already presented preliminary studies related to navigational assistance. Specifically, we expand the detection range of traversable area using RealSenseR200 [15], detect water hazards with a polarized RGB-Depth sensor [16], and detect pedestrian crosswalk [17] and crossing lights [18] at intersections. However, the road barrier is taken as a usual obstacle in these researches. In this paper, we include novel contributions and results to extend previous proof-of-concepts.

3 System Overview

Real-time image classification on mobile devices with limited computational resources often requires small memory footprint and rapid response. The critical issue is how to strike a judicious tradeoff between latency and accuracy. For this reason, we need to take the specific application scenario and the hardware platform into consideration. As shown in Figure 2, our system consists of a pair of wearable smart glasses and a mobile processor (Kangaroo [19]). The pair of wearable smart glasses is integrated with a RGB-Depth sensor (RealSense R200) and a bone-conduction headphone. On one hand, RealSense R200 has high environmental adaptability [15] and delivers real-time RGB−Infrared image streams. The color image contains rich chromatic information while infrared camera provides more stable images than color camera during walking. On the other hand, its small size and light weight make it quite suitable to integrate into a pair of wearable glasses. As far as the feedback is concerned, the bone conduction headphone transfers the recognition results to visually impaired people. This is important as visually impaired people need to continue hearing environmental sound and the bone conducting interface allow them to hear a layer of augmented acoustic reality that is superimposed on the environmental sound.

In the way to destination, the mobile processor continually calculates the distance using GPS signals between the current point of interest (POI) and the destination which we have already marked. To detail this, when the distance is less than 20 meters, the processor starts an image classification thread. Firstly, the camera perceives color images as the input images of KrNet. Secondly, KrNet outputs the results of image classification. Lastly, the bone-conduction headphone transfers sound to the visually impaired. Specifically, If the current images are classified as “barrier”, the system will remind the visually impaired of reaching the destination. When the distance is larger than 20 meters, the processor terminates the image classification thread.

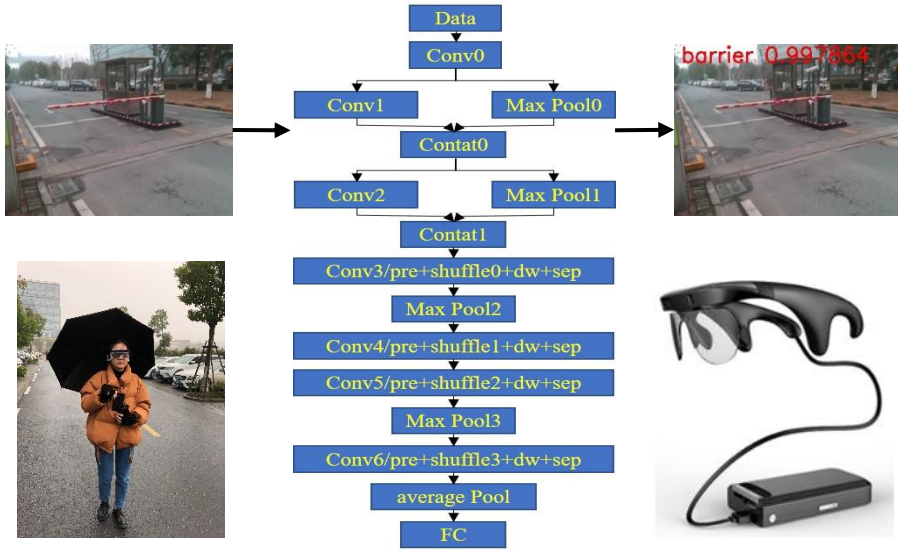

Fig. 2. Overview of the navigation system: (a) The wearable prototype. (b)The outline of the KrNet architecture from the input to the prediction. (c) The navigational assistance system consists of a pair of wearable smart glasses and a mobile processor.

4 Network Architecture

The architecture of KrNet is depicted in Table 1. The data layer takes as input a 160×160×3 image. The entry module consists of a standard convolutional layer and two Inception blocks proposed by [20], which reduces the grid-size of feature maps while expands the filter banks. The convolutional layer (Conv1) and max pooling layer (Max Pool0) take the output of the first convolutional layer (Conv0) as their input concurrently and filter it with 16 kernels of size 3×3. Afterwards, the output of Conv1 and Max Pool0 are concatenated as input of the next Inception block.

Most networks use standard convolutional layers for convolution operations where every filter operates on all of input channels until Xception [21] assumes that cross-channel correlations and spatial correlations can be mapped completely separately. Xception comes up with depthwise separable convolution which replaces a full convolutional operator with a factorized version that splits convolution into two separate layers. However, it impedes information flow between different channels, which might result in the degradation of an individual convolutional filter and weaken the representation of the network. To avoid this situation, the middle of our model consists of four depthwise separable convolutional blocks whose groups are set at 4, which require less computation than standard convolutions as well as sacrifices only a small reduction in accuracy.

As shown in Figure 3, a depthwise separable block includes a preparation pointwise convolutional layer, a shuffle layer, a depthwise convolutional layer and a separable pointwise convolutional layer. The preparation pointwise convolutional layer (GConv1) takes as input feature maps from the previous layer, reducing the channel dimension of feature maps as well as dividing the channels into 4 groups. The shuffle layer [14] reshapes, transposes and flattens the output channels to make sure that input and output channels are fully related to each other when the depthwise convolutional layer (GConv2) takes data from different groups after GConv1. Finally, the separable pointwise convolutional layer (GConv3) is used to recover the channel dimension.

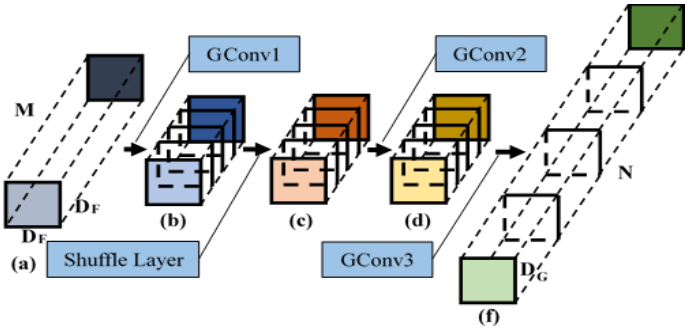

Fig. 3. (a) Input $D_F \times D_F \times M$ feature maps from the previous layer. (b) $D_F \times D_F \times 1/4N$ feature maps from GConv1. (c) $D_F \times D_F \times 1/4N$ feature maps after channel shuffle operation. (d) $D_F \times D_F \times 1/4N$ feature maps from GConv2. (f) Output $D_G \times D_G \times N$ feature maps from GConv3.

If standard convolution takes as input $h_i \times w_i \times d_i$ feature maps and applies convolutional kernel of size $k \times k \times d_j$ to produce $h_j \times w_j \times d_j$ output maps. Standard convolutional layer has the computational cost:

$$h_i \cdot w_i \cdot k \cdot k \cdot d_i \cdot d_j \tag{1}$$

Depthwise separable convolution block with shuffle operation of group 4, while each filter operates only on the corresponding input channels within same group, works almost well as regular convolutions but only cost:

$$h_i \cdot w_i \cdot d_i \cdot d_j \cdot \frac{1}{4} + h_i \cdot w_i \cdot k \cdot k \cdot d_j \cdot \frac{1}{4} + h_i \cdot w_i \cdot d_j \cdot \frac{1}{4} \cdot d_j \tag{2}$$

Our network has a computational cost of 200 million multiply-adds which is much lower than MobileNet [14]. In this regard, the efficiency is guaranteed to enable real-time inference on mobile devices.

All layers are followed by a batch normalization and a ReLU nonlinear activation function. The final average pooling reduces the spatial resolution to 1 before the Softmax loss layer. Our model was trained in Caffe CNN framework with stochastic gradient descent [22]. The dataset is available at [23]. Most of images are manual collected and resized to 320 × 240. To make the model robust against the varied types of images

from real world, we perform a group of data augmentation including horizontally flipping, adding Gaussian noise and color jittering.

Table 1. The outline of the proposed network architecture

Layer name	Type	Computational cost
Conv0	Standard convolution	$160^2\times3^2\times3\times16$
Max Pool0	Max pooling	$160^2\times3^2\times16$
Conv1	Standard convolution	$160^2\times3^2\times16\times16$
Max Pool1	Max pooling	$80^2\times3^2\times32$
Conv2	Standard convolution	$80^2\times3^2\times32\times32$
Conv3 block	depthwise convolution block with shuffle operation	$40^2\times1^2\times64\times48 + 40^2\times3^2\times48 + 40^2\times1^2\times48\times192$
Max Pool2	Max pooling	$40^2\times3^2\times192$
Conv4 block	depthwise convolution block with shuffle operation	$20^2\times1^2\times192\times72 + 20^2\times3^2\times72 + 20^2\times1^2\times72\times288$
Conv5 block	depthwise convolution block with shuffle operation	$18^2\times1^2\times288\times96 + 18^2\times3^2\times96 + 18^2\times1^2\times96\times384$
Max Pool3	Max pooling	$18^2\times3^2\times384$
Conv6 block	depthwise convolution block with shuffle operation	$9^2\times1^2\times384\times120 + 9^2\times3^2\times120 + 9^2\times1^2\times120\times480$

5 Experiments

We perform a set of experiments to validate the accuracy and reliability of KrNet. Table 2 shows the experimental results about the classification performance, which are qualified and satisfactory for the recognition of road barrier. In a binary classification task, true positive (TP) denotes the number of positive samples which were correctly predicted as positive. True negative (TN) denotes the number of negative samples that were correctly predicted as negative. False positive (FP) denotes the number of negative samples which were mislabeled as positive. And false negative (FN) denotes the number of positive samples that were mislabeled as negative. The accuracy for a class is the number of correctly labeled samples divided by the total number of samples as equation (3). The precision for a class is the number of true positives divided by the total number of samples labeled as belonging to the positives class (i.e. the sum of true positives and false positives) as equation (4).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{4}$$

Because it might make the blind confused if misclassification happens frequently. In our work, to provide navigational assistance for people with visually impairment, more attention is paid to the accuracy and precision instead of recall rate. In real-world assistance, we calculate the weighted average classification confidence of multi-frames as the final classification confidence. The current frame has the largest weight and the weight decreases in previous fames. Besides, only if the weighted average confidence is larger than confidence threshold value which is set at 0.98, will the classification result be transferred to the visually impaired. It has been proved that classification performance with weighted average confidence is better than without weighted average confidence both in accuracy and precision. Our method enables a deep-learning-based system to execute at 10-35 fps on CPU and achieves kinetic real-time scene classification.

Table 2. Experimental results.

Model type	Without weighted average confidence		With weighted average confidence		Speed on portable PC	Speed on CPU i5-7400
	Accuracy	Precision	Accuracy	Precision		
KrNet	0.9832	0.8114	0.9949	1.000	10fps	35fps

6 Conclusion and Future Work

According to the demands of people with visually impairment, we come up with a novel CNN named KrNet. The experiments demonstrate the proposed model is effective and efficient. Future works will involve in-depth experiments regarding other scenarios, such as curbs and stairs that people with visually impairment come cross in their daily life. Moreover, the proposed KrNet will serve as whole image descriptor to extract features for visual place recognition to achieve real-time place location on mobile devices.

References

1. Bourne, R.R.A., Flaxman, S.R.: Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis. Lancet Glob. Heal. 5, e888–e897 (2017).
2. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet Classification with Deep Convolutional Neural Networks. Adv. Neural Inf. Process. Syst. 1–9 (2012).
3. Shelhamer, E., Long, J., Darrell, T.: Fully Convolutional Networks for Semantic Segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39, 640–651 (2017).
4. Lin, J., Wang, W.J., Huang, S.K., Chen, H.C.: Learning based semantic segmentation for robot navigation in outdoor environment. In: 2017 Joint 17th World Congress of International Fuzzy Systems Association and 9th International Conference on Soft Computing and Intelligent Systems (IFSA-SCIS). pp. 1–5 (2017).
5. Arroyo, R., Alcantarilla, P.F., Bergasa, L.M., Romera, E.: Fusion and binarization of CNN features for robust topological localization across seasons. In: IEEE International Conference on Intelligent Robots and Systems. pp. 4656–4663 (2016).

6. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. Int. J. Comput. Vis. 115, 211–252 (2015).
7. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. ImageNet Chall. 1–10 (2014).
8. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 7-12-NaN-2015, 1–9 (2015).
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016).
10. Han, S., Mao, H., Dally, W.J.: A Deep Neural Network Compression Pipeline: Pruning, Quantization, Huffman Encoding. arXiv1510.00149 [cs]. 13 (2015).
11. Hinton, G., Vinyals, O., Dean, J.: Distilling the Knowledge in a Neural Network. Comput. Sci. 1–9 (2015).
12. Iandola, F.N., Moskewicz, M.W., Ashraf, K., Han, S., Dally, W.J., Keutzer, K.: SqueezeNet. arXiv. 1–5 (2016).
13. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. ArXiv. 9 (2017).
14. Zhang, X., Zhou, X., Lin, M., Sun, J.: ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. arXiv. 1–10 (2017).
15. Yang, K., Wang, K., Hu, W., Bai, J.: Expanding the detection of traversable area with RealSense for the visually impaired. Sensors (Switzerland). 16, (2016).
16. Yang, K., Wang, K., Cheng, R., Hu, W., Huang, X., Bai, J.: Detecting traversable area and water hazards for the visually impaired with a pRGB-D sensor. Sensors (Switzerland). 17, (2017).
17. Cheng, R., Wang, K., Yang, K., Long, N., Hu, W.: Crosswalk navigation for people with visual impairments on a wearable device. J. Electron. Imaging. 26, 1 (2017).
18. Cheng, R., Wang, K., Yang, K., Long, N., Bai, J., Liu, D.: Real-time pedestrian crossing lights detection algorithm for the visually impaired, (2017).
19. Kangaroo, http://www.kangaroo.cc/kangaroo-mobile-desktop-pro.
20. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the Inception Architecture for Computer Vision. (2015).
21. Chollet, F.: Xception: Deep Learning with Separable Convolutions. arXiv Prepr. arXiv1610.02357. 1–14 (2016).
22. Kingma, D.P., Ba, J.L.: Adam: a Method for Stochastic Optimization. Int. Conf. Learn. Represent. 2015. 1–15 (2015).
23. Road barrier dataset, http://www.wangkaiwei.org.