The Gesture Detection Algorithm Based on 3-DCGAN Range Estimation in FMCW Radar System

Xiuqian Jia(✉), Yong Wang, Mu Zhou, and Zengshan Tian

Chongqing Key Lab of Mobile Communications Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
1506894146@qq.com

Abstract. Recently, hand gesture detection has gradually become a research hotspot. We propose a Region-based Faster Convolutional Neural Network (F-RCNN) gesture detection method based on Frequency Modulated Continuous Wave (FMCW) radar using 3-Dimensions Deep Convolutional Generative Adversarial Networks (3-DCGAN). Specifically, this paper adopts FMCW radar for hand gesture data acquisition, and estimates the distance of the hand gesture using the regularity of the change of echo frequency and emission frequency of radar signals. Then the semantic label maps of the generated images of distance are sent to the 3-DCGAN to extend datasets. After that, the original images and the images generated by the 3-DCGAN are simultaneously sent to F-RCNN for training. The results show that the proposed approach increases the mAP by 3% compared to the baseline F-RCNN. Besides, the proposed method not only effectively solves the problem of small amount of hand gesture data, but also the manpower and material resources consumed by collecting data.

Keywords: F-RCNN · FMCW radar · Gesture detection

1 Introduction

With the rapid development of computer radio in 5 Generation [1–3] and human-computer interaction, the use of gesture detection methods to replace traditional mechanical keyboards and mice has gradually become a research hotspot [4–7]. The traditional gesture detection method is mainly based on the processing method of the camera [4, 5]. After the scene is taken by the camera, a static image is obtained, and then the image content is detected by a computer graphics algorithm [4]. The camera method is divided into an optical sensor method and a depth sensor method [6]. Although the optical sensor camera method can provide accurate detection, it is susceptible to illumination conditions [6, 7], and can't effectively provide accurate depth information, and the optical camera method has a great trouble for people's privacy. For depth sensor cameras, performance is greatly reduced during outdoor measurements due to sun damage. Compared with the camera method, the gesture detection method based on radar can effectively overcome the above shortcomings and has been widely used [8–11]. Using radar for gesture detection, the hand is modeled as a single

M. Jia et al. (Eds.): WiSATS 2019, LNICST 280, pp. 242–250, 2019.
https://doi.org/10.1007/978-3-030-19153-5_25

hard object [8, 10], in fact, it can scatter the signal from the radar. The Frequency Modulated Continuous Wave (FMCW) radar can perceive the hand as a multi-scattering object and captures its local microscopic motion [12–14]. Therefore, the hand is modeled as a non-rigid object. In the context of dynamic gesture detection, gestures produce multiple reflections from different parts of the hand, with different ranges values vary over time. In [15], the FMCW radar is used to estimate the distance-Doppler map of the radar to detect the gesture signal, and the 3-Dimension (3D) spatial position of the gesture is estimated by the radar. Since the frequency difference is determined by the frequency aspect ratio of the FMCW signal, in order to accurately estimate the distance information of the gesture, a highly linear frequency aspect ratio can be adopted to improve the ranging accuracy [16]. Besides, the Region-based Faster Convolutional Neural Network (F-RCNN) method [17] selects the feature candidate boxes in the picture and uses the classifier to discriminate, and the gesture detection accuracy is improved.

Based on the above analysis, this paper proposes a gesture detection method for 3-Dimensions Deep Convolutional Generative Adversarial Networks (3-DCGAN) range estimation based on FMCW radar. Firstly, the FMCW radar is used for gesture data acquisition. According to the law of the echo frequency and the transmission frequency of the radar signal, we can calculate the distance of the gesture. Secondly, we send the semantic label maps of the Range-Time-Map (RTM) images into the 3-DCGAN network model to extend the dataset. Thirdly, the original range images and the images generated by the 3-DCGAN model are simultaneously sent to F-RCNN for jointly training. We adopt 3-DCGAN model in our paper, and 3-DCGAN extracts both high resolution and low resolution features to improve the dimensional information of the original image. Since high resolution and low resolution features are sensitive in the process of feature extraction, the detection performance of the generated images of 3-DCGAN will better. Besides, the 3-DCGAN model are used to extend the datasets, the experimental results show that the proposed method not only effectively solves the problem of small amount of hand gesture data, but also the manpower and material resources consumed by collecting data, and greatly improves the accuracy of hand gesture detection.

2 FMCW Radar

This paper adopts FMCW radar sensor for hand gesture data acquisition. FMCW radar transmits high-frequency continuous signal, and the frequency of the transmitted triangle signal changes linearly with time. The echo frequency of the radar signal has the same regularity as the change of the transmission frequency, and they only differ by one time difference. The FMCW radar prototype used for hand gesture data acquisition is shown in Fig. 1.

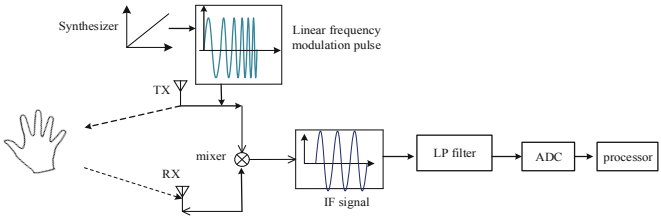

Fig. 1. The process of acquiring gesture data by FMCW radar.

It is observed from Fig. 1 that FMCW radar generates a chirp signal (linear frequency modulation pulse) through a synthesizer, and transmitted signal encounters an obstacle (such as a hand) for reflection, and the receiving antenna (RX) captures a reflected signal. With a mixer, the differential signal (intermediate frequency (IF)) of the transmitted and received signals is obtained. Then the IF signal passes through a low-pass filter and is digitized by the analog-to-digital converter (ADC). The ADC sample rate is commensurate with the largest range. Then the digital data is for Fast Fourier Transform (FFT) processing.

In general, the frequency of the transmitted signal can be expressed as:

$$f_T(t) = f_0 - B/2 + Bt/T \tag{1}$$

where f_0 is the center frequency of the transmitted signal, T is the signal period, and B is the bandwidth of the transmitted signal, then the transmitted signal can be expressed as

$$\begin{aligned} V_T(t) &= A_T \cos\left(2\pi \int_0^t f_T(t)dt\right) \\ &= A_T \cos\left(2\pi(f_0 - \frac{B}{2})t + \frac{\pi B}{T}t^2\right) \end{aligned} \tag{2}$$

where A_T is the amplitude of the transmitted signal. Assuming that the range from the radar to the gesture is R, the delay of the received signal obtained by the radar relative to the transmitted signal is t_d, $\Delta\varphi$ is the Doppler phase shift, and c is the speed of light, then $t_d = 2R/c$. So the received signal can be expressed as

$$V_R(t) = A_R \cos\left(2\pi(f_0 - \frac{B}{2})(t - t_d) + \frac{\pi B}{T}(t - t_d)^2 + \Delta\varphi\right) \tag{3}$$

where A_R is the amplitude of the received signal. When the transmitted signal is mixed with the received signal and passed through a low-pass filter, the IF signal is obtained

$$S(t) = \frac{A_R A_T}{4} \cos\left(2\pi t_d(f_0 - \frac{B}{2})\right) \cos\left(\frac{\pi B}{T}t_d^2 - \frac{2\pi B}{T}t_d t + \Delta\varphi\right) \tag{4}$$

The frequency of the IF signal is

$$f_{IF} = Bt_d/T \tag{5}$$

The correspondence relationship between the estimated range R of the gesture and the frequency f_{IF} of the IF signal can express as

$$R = \frac{cT}{2B} f_{IF} \tag{6}$$

where B/T is the slope of the chirp signal.

This paper mainly analyzes the fast time domain (one sweep time) of the FMCW radar signal. In the fast time domain, according to the FFT analysis of the frequency sweep signal, the spectrum of the IF signal can be obtained, and then the frequency point corresponding to the gesture target is obtained according to the spectral peak search.

The parameter information of one frame (128 sweep time) of data can be obtained by using the distance estimation methods of the gesture. Since the time of one frame of data is only 40 ms, the change of the gesture target is almost negligible in one frame. This paper has set the observation duration to 32 frames by some experiments. Therefore, we use the FMCW radar equipment to analyze the obtained data as described above, and the distance of the gesture signal can be obtained respectively, as shown in Fig. 2.

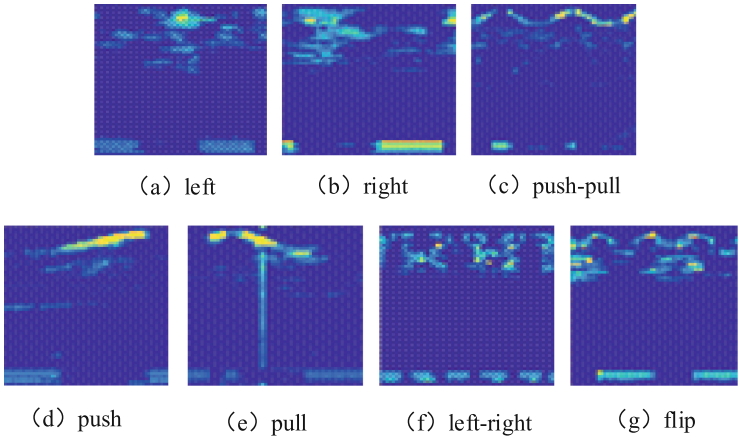

Fig. 2. Gesture distance signal diagram

3 Model Description

3.1 3-DCGAN Model

In our paper, we adopt a 3-Dimension generator, a multi-scale discriminator architecture and a robust adversarial learning objective function, named 3-DCGAN.

Because 3-DCGAN can extract both high resolution and low resolution features, and they are sensitive in the process of feature extraction. Now we will introduce it in detail.

Firstly, we collect the data through the FMCW radar board, and the data is pre-processed to obtain range, speed, and angle maps of different gestures. Then we get the semantic label maps of the three kinds of gestures. We describe our generator into four sub-networks: G1, G2, G3 and G4. In our model, G1, G2, G3 are represented the high-resolution, low-resolution and the original feature maps, respectively. And G4 represents the residual blocks. G1 and G2 are extracted from the original images, and the residual network G3 is trained on the original images. Then the different dimensional features G1, G2 and the original image features G3 are merged. Finally, the fusion feature is input into the residual network G4 for image generation. The resolution of image G1 is 4 times the previous output size G1 (2× along each image dimension) and the resolution of image G2 is 1/4 times of the previous output size G1.

In our paper, our original generator G3 is adopted the model by Johnson et al. [5], which includes a convolutional front-end G_3^F, a set of residual blocks G_3^R [6], and a transposed convolutional back-end G_3^B. G_1 and G_2 can only include a convolutional front-end G_1^F and G_2^F G_4 is composed of a set of residual blocks G_4^R and a transposed convolutional back-end G_4^B. In our model, G_4^R is different from G_3, which is the element-wise sum of three feature maps: the output feature map of G_1^F, the output feature map of G_2^F, and the last feature map of the G_3^B.

In order to improve the resolutions and to extract gesture image features effectively. During training, we train G_3 generator firstly. Secondly, we train the remaining network structure follow a sequence. Thirdly, we jointly fine-tune all the networks together. This idea can be found in [18, 19] and conditional image generation [20, 21]. As for discriminator, we adopt multi-scale discriminators in [22]. The 3-DCGAN generator description can be observed in Fig. 3. Among them, black square dotted line represents residual blocks.

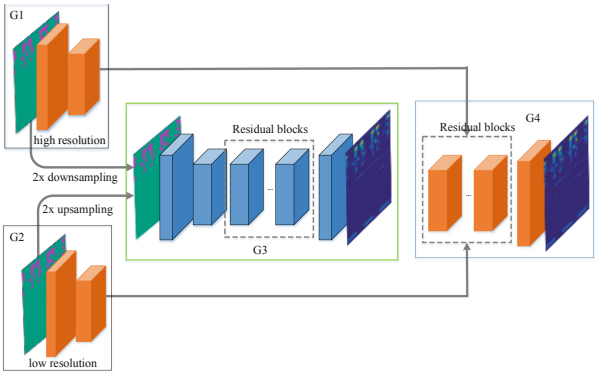

Fig. 3. 3-DCGAN generator network structure.

3.2 F-RCNN Model

The F-RCNN consists of two modules: the region proposal network (RPN) candidate extraction module and Fast R-CNN detection module. The RPN is a full convolutional neural network that is adopted for extracting candidate regions. Then, Fast R-CNN detects and identifies targets based on the extracted RPN proposal. To infer the hand gesture location, the more advanced hand gesture detection network needs to use the regional recommendation algorithm. Although SPP-net [23] and Fast R-CNN network have reduced the detection time, the calculation of area recommendation is still time-consuming. As a result, RPN is proposed to extract the interest area. Since RPN shares the convolution feature of the whole image with the entire detection network, the area recommendation time is largely reduced. The F-RCNN diagram for hand gesture detection is shown in Fig. 4.

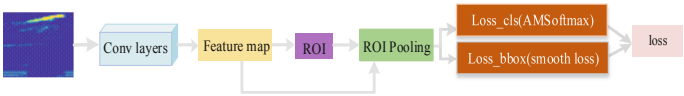

Fig. 4. The F-RCNN based hand gesture detection diagram

4 Experiment Results

4.1 mAP Performance Comparison

In this paper, the initial baseline mAP of F-RCNN is 69.5%. When adopting the 3-DCGAN model to generate images, the mAP rises to 72.9%. In addition, the performance of left-right gesture detection is best, and the mAP is 85.9%. This is because when the radar captures the gesture data with the hand slides left-right, the distance information of the gesture are processed and the image features are obvious, as shown in Fig. 5 and Table 1.

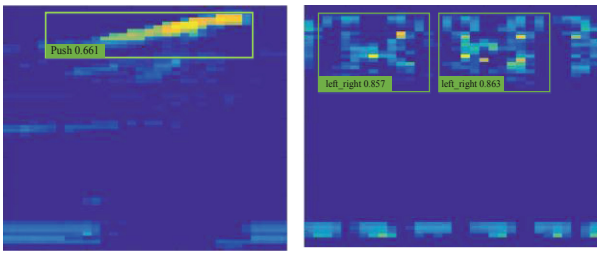

Fig. 5. The detection effect in F-RCNN model.

Table 1. Detection effect for different gestures (%).

Method	Avg	Left	Right	Push-pull	Push	Pull	Left-right	Flip
F-RCNN [14]	69.5	65.5	60.7	60.1	62.9	74.5	86.2	76.7
F-RCNN + 3-DCGAN	72.9	68.4	62.9	66.5	65.8	78.2	85.9	82.9

4.2 Impact of Generated Images

In order to evaluate the proposed method effectively, Table 2 shows the results of comparing the different numbers of 3-DCGAN generated images with the original images in this paper. Among them, 0 (basel) means the number of the original image is 7000 without the 3-DCGAN generated images. Gesture-7000 means only 7000 3-DCGAN generated images are used. Gesture + 7000 represents 7000 3-DCGAN generated images plus the 7000 original images. When the original 7000 gesture images and the 7000 3-DCGAN generated images are simultaneously fed into the F-RCNN model for training, the mAP is 71.8%. Compared with the baseline, the mAP is increased by 2.3%. This shows that training with images generated by the 3-DCGAN model perform better than the original images, indicating that the effectiveness of training 3-DCGAN model.

We continue to increase the number of 3-DCGAN generated images, when the original images and the number of 2 × 3-DCGAN images are sent to F-RCNN, the mAP reaches a maximum of 72.9%. However, we furtherly increase the 3-DCGAN generated images for training, the detection accuracy is reduced. This is because the learning machine tends to assign a uniform prediction probability to all training samples when there are too many 3-DCGAN images.

Table 2. The comparison of adding different numbers of 3-DCGAN generated images and the original images.

3-DCGAN	mAP(%)
0 (basel)	69.5
Gesture − 7000	71.2
Gesture + 7000	71.8
Gesture + 15000	72.9
Gesture + 21000	72.1

5 Conclusion

In this paper, 3-DCGAN is firstly presented to generate high quality images to expand the hand gesture dataset. Then, we adopt the generated and the original images to Faster RCNN for training and testing. The results show that the proposed approach increases the mAP by 3% compared to the baseline F-RCNN. We adopt 3-DCGAN model in our paper, and 3-DCGAN extracts both high resolution and low resolution features to improve the dimensional information of the original image. Since high

resolution and low resolution features are sensitive in the process of feature extraction, the detection performance of the generated images of 3-DCGAN will better. Besides, the 3-DCGAN model are used to extend the datasets, the experimental results show that the proposed method not only effectively solves the problem of small amount of hand gesture data, but also the manpower and material resources consumed by collecting data, and greatly improves the accuracy of hand gesture detection.

References

1. Jia, M., Gu, X., Guo, Q., Xiang, W., Zhang, N.: Broadband hybrid satellite-terrestrial communication systems based on cognitive radio toward 5G. IEEE Wirel. Commun. **23**(6), 96–106 (2016)
2. Jia, M., Liu, X., Gu, X., Guo, Q.: Joint cooperative spectrum sensing and channel selection optimization for satellite communication systems based on cognitive radio. Int. J. Satell. Commun. Network. **35**(2), 139–150 (2017)
3. Jia, M., Liu, X., Yin, Z., Guo, Q., Gu, X.: Joint cooperative spectrum sensing and spectrum opportunity for satellite cluster communication networks. Ad Hoc Netw. **58**, 231–238 (2016)
4. Hjelmas, E., Low, B.K.: Face detection: a survey. Comput. Vis. Image Underst. **83**(3), 236–274 (2001)
5. Mitra, S., Acharya, T.: Gesture recognition: a survey. IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev. **37**(3), 311–324 (2007)
6. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: a reference survey. ACM Comput. Surv. (CSUR) **35**(4), 399–458 (2003)
7. Gavrila, D.M.: The visual analysis of human movement: a survey. Comput. Vis. Image Underst. **73**(1), 82–98 (1999)
8. Raj, B., Kalgaonkar, K., Harrison, C., Dietz, P.: Ultrasonic doppler sensing in HCI. IEEE Pervasive Comput. **11**(2), 24–29 (2012)
9. Wan, Q., Li, Y., Li, C., Pal, R.: Gesture recognition for smart home applications using portable radar sensors. In: IEEE Conference on Engineering in Medicine and Biology Society, pp. 6414–6417, August 2014
10. Molchanov, P., Gupta, S., Kim, K., Pulli, K.: Multi-sensor system for driver's hand-gesture recognition. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 1, pp. 1–8. IEEE (2015)
11. Molchanov, P., Gupta, S., Kim, K., et al.: Short-range FMCW monopulse radar for hand-gesture sensing. In: 2015 IEEE Radar Conference (RadarCon), pp. 1491–1496. IEEE (2015)
12. Molchanov, P., Gupta, S., Kim, K., et al.: Multi-sensor system for driver's hand-gesture recognition. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 1, pp. 1–8. IEEE (2015)
13. Wang, S., Song, J., Lien, J., Poupyrev, I., Hilliges, O.: Interacting with soli: exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum. In: Proceedings of the 29th Annual Symposium on User Interface Software and Technology, pp. 851–860. ACM (2016)
14. Molchanov, P., Gupta, S., Kim, K., et al.: Short-range FMCW monopulse radar for hand-gesture sensing. In: 2015 IEEE Radar Conference (RadarCon) on Automatic Face and Gesture Recognition (FG), vol. 5, pp. 1–6. IEEE (2015)
15. Piper, S.O.: Homodyne FMCW radar range resolution effects with sinusoidal nonlinearities in the frequency sweep. In: Proceedings of IEEE International Radar Conference, pp. 563–567 (1995)

16. Molchanov, P., Gupta, S., Kim, K., Pulli, K.: Multi-sensor system for driver's hand-gesture recognition. In: AFGR (2015)
17. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. (2016)
18. Denton, E., Chintala, S., Szlam, A., Fergus, R.: Deep generative image models using a Laplacian pyramid of adversarial networks. In: Advances in Neural Information Processing Systems (NIPS) (2015)
19. Huang, X., Li, Y., Poursaeed, O., Hopcroft, J., Belongie, S.: Stacked generative adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
20. Chen, Q., Koltun, V.: Photographic image synthesis with cascaded refinement networks. In: IEEE International Conference on Computer Vision (ICCV) (2017)
21. Zhang, H., et al.: StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
22. Wang, T., Liu, M., Zhu, J., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional GANs. In: European Conference on Computer Vision (ECCV) (2017)
23. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **37**, 1904–1916 (2015)