LETTER

Complex-Valued Fully Convolutional Networks for MIMO Radar Signal Segmentation

Motoko TACHIBANA[†a)], **Kohei YAMAMOTO**[†], ***and*** **Kurato MAENO**[†], ***Members***

SUMMARY Radar is expected in advanced driver-assistance systems for environmentally robust measurements. In this paper, we propose a novel radar signal segmentation method by using a complex-valued fully convolutional network (CvFCN) that comprises complex-valued layers, real-valued layers, and a bidirectional conversion layer between them. We also propose an efficient automatic annotation system for dataset generation. We apply the CvFCN to two-dimensional (2D) complex-valued radar signal maps (r-maps) that comprise angle and distance axes. An r-maps is a 2D complex-valued matrix that is generated from raw radar signals by 2D Fourier transformation. We annotate the r-maps automatically using LiDAR measurements. In our experiment, we semantically segment r-map signals into pedestrian and background regions, achieving accuracy of 99.7% for the background and 96.2% for pedestrians.

key words: *radar signal segmentation, fully convolutional network, complex-valued network, semantic segmentation, MIMO radar*

1. Introduction

Radar is expected to play an important role in advanced driver-assistance systems because of its robustness to various environmental factors such as lightning and weather conditions [1]. Frequency-modulated continuous wave (FMCW) radar can measure distances by analyzing the frequency shifts of beat signals. Combining it with a multiple-input multiple-output (MIMO) antenna array, FMCW radar can also measure angles by analyzing the phase shifts among multi-channel beat signals. As such, FMCW MIMO radar gives a two-dimensional (2D) radar signal map (r-map) as a complex-valued matrix whose rows indicate angles and whose columns indicate distances.

On another front, significant progress has been made in image recognition by deep convolutional networks. Long et al. proposed the fully convolutional network (FCN), a semantic segmentation technique that classifies objects in an image by pixel-wise segmentation [2].

In recent years, a great deal of research on the semantic segmentation of images has been conducted, whereas far fewer studies have focused on radar signals. In this paper, we propose a novel semantic segmentation method for r-maps by using a CvFCN. The inputs to a CvFCN are complex values, but the outputs must be real values for a classification task. Therefore, we have designed an architecture that transfers complex-valued and real-valued layers bidirectionally. This allows a CvFCN to be trained as a segmentation model end to end.

Generally, in order to train deep neural networks, it is necessary to prepare a large amount of annotated data. For image recognition, we can obtain semantically segmented annotated data from an existing database such as the Pascal Visual Object Classes dataset. However, as far as we know, such annotated data for radar signals have not been reported. Therefore, we began by preparing many annotations for the semantic segmentation of r-maps. These annotations are simpler in shape compared with those for images; however, it is difficult even for a person to segment element-wise object classes on r-maps because they include pseudo reflections, called ghosts. Ghosts are generated from side lobes of Fourier transforms or multi-path reflections. Thus, in the present study, we created annotations automatically by referencing light detection and ranging (LiDAR) data and the r-map intensity distributions.

2. Automatic Annotation System

2.1 Conversion of Radar Signal to R-Map

Before generating the annotations, we convert the radar beat signals into an r-map by 2D fast Fourier transform (FFT) as follows:

$$B_t(d,\theta) = \frac{1}{NM}\sum_{r=0}^{M-1}\sum_{\tau=0}^{N-1} q_r b_{t-N+\tau+1,r} e^{-i\,2\pi\left(\frac{\tau d}{N}+\frac{r\theta}{M}\right)}, \qquad (1)$$

$$q_r = \begin{cases} 1, & 0 \le r \le K \\ 0, & K < r < M \end{cases},$$

where $b_{t,r}$ is the complex-valued beat signal of aligned receiving antenna r at sampling time t, K is the number of receiving antennas, and M and N are the FFT window sizes in the antenna-alignment and time directions, respectively. Term τ is the index for the beat signal in the FFT window in the time direction, and $B_t(d,\theta)$ is the complex-valued element of the r-map at distance d and angle θ.

Figure 1 (a) and (b) show an image taken from the radar position and the intensity of its r-map, respectively. As shown in Fig. 1 (b), the region of strongest intensity near the center of the r-map indicates where the person is. Although there are other regions of strong intensity, there are no targets. Such strong intensity regions without targets are called ghosts. Note that to ensure stationary objects are ignored, we apply a moving-target-indication filter when we

Manuscript received September 29, 2017.
Manuscript revised January 5, 2018.
Manuscript publicized February 20, 2018.

[†]The authors are with Oki Electric Industry Co., Ltd., Warabishi, 335–8510 Japan.
a) E-mail: tachibana233@oki.com

DOI: 10.1587/transinf.2017EDL8214

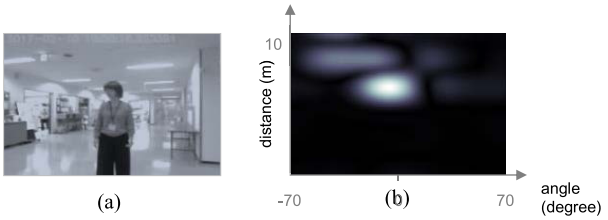


Fig. 1 (a) Image taken from radar position. (b) Intensity of corresponding r-map.

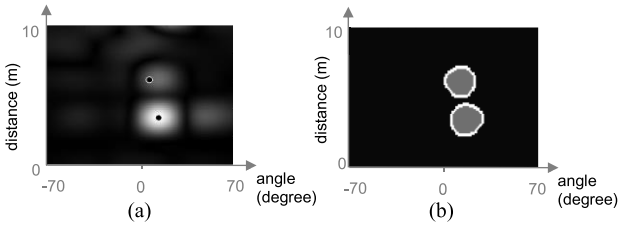


Fig. 2 Example of automatic annotation.

generate r-maps in our system.

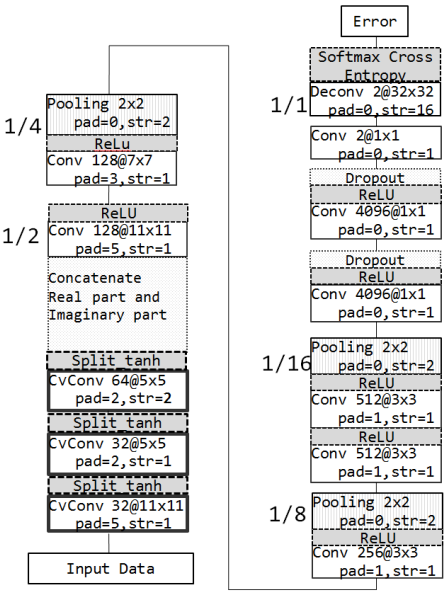
2.2 Automatic Annotation

Simultaneously with our radar, we used 2D LiDAR at almost the same position. We extract the pedestrian regions in r-maps automatically by expanding from those points on the r-map at which the LiDAR detects pedestrians. An example of this process is shown in Fig. 2. Figure 2 (a) shows an r-map overlaid with two points at which the LiDAR detected pedestrians, and Fig. 2 (b) shows the corresponding automatically created annotation; gray regions indicate pedestrians, black regions are background, and white regions are not used for training. This process can automatically prepare enough annotation maps to train a deep neural network model. In order to improve the angular resolution of the annotation, we use the Khatri–Rao product [3] when creating r-maps.

3. Complex-Valued Fully Convolutional Network Model

In recent work on complex-valued networks, Guberman [4] showed that fully complex-valued convolutional networks can detect meaningful phase structures in complex-valued data extracted from an image using the Sobel kernel. Motivated by that work, we designed a partially complex-valued FCN as a CvFCN.

To segment an r-map semantically, the output and input must comprise real and complex values, respectively. The proposed CvFCN was designed to learn from complex-valued input to real-valued output end to end. We describe here the CvFCN concatenation process that converts bidirectionally between a complex-valued layer and a real-valued layer. Figure 3 shows in detail the CvFCN used in our experiment described in Sect. 4. As shown in Fig. 3, the r-map is applied to the three complex-valued convolution layers and converted to 64 channels; each of which is a complex-valued matrix. By dividing them into real and imaginary parts, we generated two sets of 64 real-valued matrices and used them as 128-channel inputs to the subsequent layers. With this architecture, it is possible to learn the r-map topology end to end.



Notes:
CvConv: complex-valued convolution
Conv: real-valued convolution
Deconv: deconvolution
Split_tanh: activation function for complex values
ReLU: activation function for real values

"Conv 32@5x5, pad=2, str=1" means convolution with 32 channels, 5×5 kernel size, 2 for spatial padding size, and 1 for stride width of filter applications.

Fig. 3 Configuration of proposed CvFCN.

Because the phase distributions of an r-map contain finer information such as the fractional movements of people or finer distance changes, we believe that it is important to use not only intensity but also phase structure as input data for restraining intensity noise when segmenting radar signals. We follow the forward and backward propagation scheme of Ref. [5] when dealing with the complex-valued networks.

As shown in Fig. 3, we use a split hyperbolic tangent function for activation following all of the complex convolutional layers. It activates the real and imaginary parts separately as follows:

$$f(z) = f(x + iy) = \tanh x + i \tanh y. \quad (2)$$

We halve the size in the distance–angle direction by setting a stride of two at the third layer of the complex convolution. We use a rectified linear unit (ReLU) as the activation function for the outputs of the real convolutional layers, and MaxPooling to compress information in the distance–angle direction. We repeat this process until the

r-map is reduced in size by a factor of 1/16. Finally, we expand the feature maps so generated into the same size of input r-map to obtain the probability of each pixel being either pedestrian or background.

4. Experiment and Evaluation

We acquired indoor data by using a 24-GHz FMCW MIMO radar where there were some pedestrians most of the time. For learning, we used 27,318 data that were obtained in three hours on two days. For evaluation, in order to ensure independence, we used 503 data that were acquired on another day. All data were acquired at one location; however, we set up the measurement hardware (i.e., the radar and LiDAR) on each collection day. Thus, the installation conditions may have differed slightly according to the day.

To train the CvFCN, we set the batch size to 100, whereas the training batch size in Ref. [2] was 1. Because segmenting an r-map semantically gives simpler shapes than doing so for an image, we expected that a larger batch size would accelerate the convergence of learning. We also used Adam [6] for optimization, the parameters of which were 0.0001 for alpha, 0.8 for beta1, and 0.9 for beta2.

We evaluated the accuracy of signal segmentation as follows, using the same metric as that in Ref. [2]:

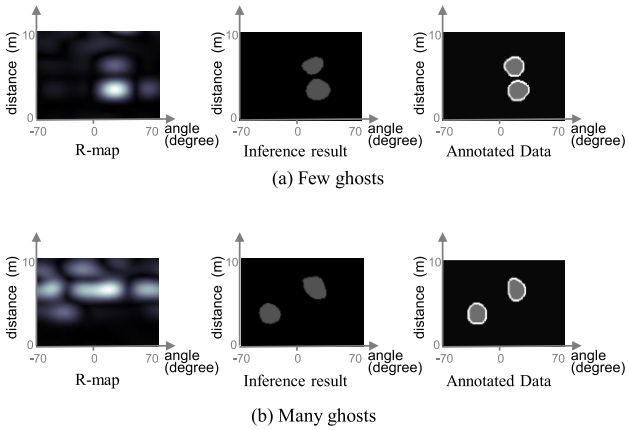
- Pixel accuracy: n_{ii}/t_i, (3)
- Mean IU: $(1/n_{cl}) \sum_i^{n_{cl}} \left\{ n_{ii} \middle/ \left(t_i + \sum_j n_{ji} - n_{ii} \right) \right\}$,

where i is the class index to be distinguished, t_i is the number of pixels of class i, n_{ij} is the number of pixels of class i predicted to belong to class j, and n_{cl} is the number of classes.

Figure 4 shows two examples of the results. From left to right, the columns in Fig. 4 show the r-map input data, the inferential data, and the ground truth. The gray regions of the inferential and ground truth indicate pedestrians. Figure 4 (b) shows that the CvFCN can detect pedestrian regions without false detection of ghosts, which are high-intensity regions in the r-map.

Figure 5 shows the transitions corresponding to the learning iteration number of an error and the resulting accuracy, which is defined by Eq. (3). As a result, we achieved recognition accuracies of 99.7% for the background and 96.2% for pedestrians, as shown in Fig. 5 (b).

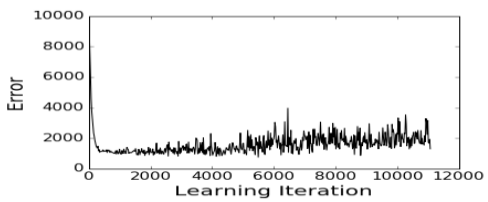
In Fig. 5 (a), the error tends to increase proportionally to iteration number. This suggests that the probabilities of belonging to each class are becoming close. By reducing the learning rate according to the number of iterations, it might be possible to suppress the error growth. In addition, the pedestrian pixel accuracy decreases in Fig. 5 (b). At an early-stage iteration, the inferred pedestrian regions could be wider and even include background pixels. Such false-detection errors are evaluated in Mean IU of Eq. (3), since the denominator includes the term $\sum_j n_{ji} - n_{ii}$. As shown in Fig. 5 (c), Mean IU improves as the number of iterations increases. This indicates that segmentation performance



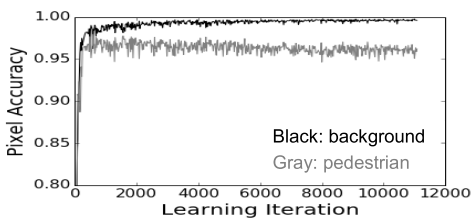
(a) Few ghosts

(b) Many ghosts

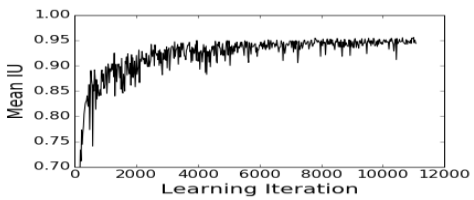
Fig. 4 Two examples of inference results.



(a) Error



(b) Pixel accuracy



(c) Mean IU

Fig. 5 Evolution of evaluation values: (a) error; (b) pixel accuracy; (c) mean IU.

could be improved iteratively. We suppose the decrease in pedestrian pixel accuracy is attributable to the slightly wider annotation regions of pedestrian, because boundaries of targets on r-maps blur. We expect that pedestrian pixel accuracy can be improved by adjusting the method for setting annotation boundaries.

5. Discussion and Conclusion

In the present study, we proposed a radar signal segmentation method using a CvFCN and an automatic annotation system for dataset generation. Our unique approach of using a bidirectional conversion layer between complex-valued

and real-valued layers achieved backward propagations for complex-valued parameters from real-valued teaching signals. Because the CvFCN is able to learn phase properties in r-maps end to end, it is effective at restraining ghosts. In addition, our automatic annotation system makes it possible to prepare many annotation maps without much manual effort for supervised learning.

In future work, we will acquire data on various objects and at various places to enhance functionality and improve accuracy. We will also compare the CvFCN with a real-valued FCN to verify the effectiveness of the CvFCN.

References

[1] S.H. Jeong, J.E. Lee, S.U. Choi, J.N. Oh, and K.H. Lee, "Technology analysis and low-cost design of automotive radar for adaptive cruise control system," International Journal of Automotive Technology, vol.13, no.7, pp.1133–1140, 2012.

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.3431–3440, 2015.

[3] Y. Wakamatsu, H. Yamada, and Y. Yamaguchi, "MIMO Doppler radar using Khatri-Rao product virtual array for human location estimation," Electromagnetics of the 2014 IEEE International Workshop on, pp.30–31, 2014.

[4] N. Guberman, "On Complex Valued Convolutional Neural Networks," arXiv preprint arXiv:1602.09046, 2016.

[5] A. Hirose, "Complex-valued neural networks: Advances and applications," John Wiley & Sons, 2013.

[6] D.P. Kingama and J.L. Ba, "Adam: A Method for Stochastic Optimization," International Conference on Learning Representations (ICLR) 2015, Main Conference Poster Session, no.11, 2015.