引用格式：LIANG Guang-yu，CHENG Liang-lun，HUANG Guo-heng，*et al*. Object Detection of Millimeter-wave Image Based on Spatial-transformer RCNN with Deblurring[J]. *Acta Photonica Sinica*，2020，**49**(2)：0210004
梁广宇，程良伦，黄国恒，等.基于去模糊空间变换 RCNN 的毫米波图像目标检测[J].光子学报，2020，**49**(2)：0210004

基于去模糊空间变换 RCNN 的毫米波图像目标检测

梁广宇，程良伦，黄国恒，徐利民
（广东工业大学 计算机学院，广州 510000）

摘　要：提出一种包含去模糊的空间变换区域卷积神经网络的目标检测算法.首先，基于主动毫米波圆柱扫描成像原理对人体进行三维成像（频率 24～30 GHz），建立毫米波图像数据集.然后，估计毫米波图像的模糊核，通过卷积去噪网络获得图像先验知识，将其集成到半二次分裂的优化方法中，以实现非盲目去模糊.最后，由定位网络、网格生成器和采样网络三部分组成空间变换网络，将它融入到特征提取网络中，在去模糊后实现目标检测.通过该非盲目去模糊算法得到的图像的峰值信噪比可达 27.49 dB，目标检测算法的平均精度可达 80.9%.实验结果表明，与现有的先进方法相比，该方法可以有效地提高图像质量和检测精度，为毫米波图像中隐藏危险品的目标检测提供了新的技术支持

关键词：安全检测；毫米波图像；目标检测；空间变换区域卷积神经网络；非盲目去模糊
中图分类号：TP391.4；TP18　　**文献标识码**：A　　**doi**：10.3788/gzxb20204902.0210004

Object Detection of Millimeter-wave Image Based on Spatial-transformer RCNN with Deblurring

LIANG Guang-yu，CHENG Liang-lun，HUANG Guo-heng，XU Li-min
（*School of Computer*，*Guangdong University of Technology*，*Guangzhou* 510000，*China*）

Abstract：An object detection algorithm of spatial-transformer regional convolutional neural network with deblurring was proposed. Firstly，based on the principle of active millimeter-wave cylindrical scanning imaging，the human body is three-dimensionally imaged（frequency range from 24 GHz to 30 GHz），and a millimeter wave image data set is established. Then the blur kernel of the millimeter-wave image is estimated. The image prior knowledge is obtained by the convolutional denoiser network and is integrated into an optimization method of half quadratic splitting to achieve non-blind deblurring. Finally，the spatial transform network，composed of a localization net，a grid generator，and a sampling network，is inserted into the feature extraction network to achieve object detection after deblurring. With the proposed non-blind deblurring algorithm，peak signal to noise ratio of the image can reach 27.49 dB. Mean average

Foundation item：National Key Research and Development Program of China（No.2016YFB1200402-019），Special Funds for Applied Science and Technology Research and Development in Guangdong Province（No.2015B090923004），Guangdong Provincial Key Laboratory of Cyber-Physical System（No.2016B030301008），NSFC-Joint fund of Guangdong Province（Nos.U1801263，U1701262），National High Resolution Earth Observation Major Project（No.83-Y40G33-9001-18/20）

First author：LIANG Guang-yu（1995－），male，M.S. degree，mainly focuses on image processing of millimeter-wave and deep learning. Email：lianggyu@outlook.com
Corresponding author：HUANG Guo-heng（1985－），male，lecturer，Ph. D. degree，mainly focuses on computer vision and pattern recognition. Email：kevinwong@gdut.edu.cn
Supervisor：CHENG Liang-lun（1964－），male，professor，Ph.D. degree，mainly focuses on terahertz technology，cyber-physical system and image processing. Email：llcheng@gdut.edu.cn

Received：Aug.29，2019；**Accepted**：Nov.20，2019

precision of object detection algorithm can reach 80.9%. The experimental results show that the image quality and detection accuracy can effectively be improved through the proposed method compared with some state-of-the-art methods. New technical support is provided for object detection of hidden dangerous goods in millimeter-wave images.

Key words: Security inspection; Millimeter-wave image; Object detection; Spatial-transformer regional convolutional neural network; Non-blind deblurring

OCIS Codes: 100.0100; 100.4994; 100.4996; 110.0110; 110.6880

0 Introduction

With increasing need for security inspection of transportation and public places, millimeter-wave imaging systems are developed to automatically determine and locate the corresponding categories of forbidden objects, like knives and guns, on human bodies.

As for enhancing and denoising millimeter-wave images, image processing methods such as wavelet transform, interpolation, and histogram equalization are used. In Ref.[1], four interpolation methods are investigated to improve the image quality of a Passive Millimeter-Wave (MMW) image. MATEOS J et al.[2] proposed a robust Bayesian multiframe blind image deconvolution method that approximates the posterior distribution of the blur by a Dirichlet distribution. But traditional algorithms of image enhancement are sensitive to noise. And it is easy to cause the millimeter-wave image to be too weak or excessively enhanced. In recent years, it has been made significant progress on the deblurring of the single image[3]. Deblurring with statistical properties of a particular domain is used in many recent methods, such as text[4], face[5] and low-light image[6]. The Maximum A Posteriori (MAP) estimation is used in non-blind deblurring of the current work, with differences in the type of the image prior knowledge they employ[7]. Instead of learning the discriminant model of MAP estimation, a simple Convolutional Neural Network (CNN) is used to learn the denoiser. In this work, a deblurring algorithm for millimeter-wave images based on the model optimization method is proposed. The model-based optimization method can handle the task of image restoration by specifying the degradation matrix.

At present, the object detection of hidden objects in millimeter-wave image can be roughly divided into two categories: traditional methods and deep learning methods. The traditional methods are mostly used which are generally based on the idea of image classification. XIAO Z L et al.[8] proposed a Passive Millimeter-Wave (PMMW) image contraband detection method based on Haar-like features, and the AdaBoost algorithm is adopted to get the strong classifier. LI Z et al.[9] choose saliency, Scale Invariant Feature Transform (SIFT) and Histogram of Oriented Gradient (HOG) feature to form image descriptors, combined with linear SVM for object/non-object classification. In recent years, deep learning has been applied to object detection of the millimeter-wave image. LIU T et al. [10] investigated the deep learning-based framework, Faster R-CNN, in Active Millimeter-Wave (AMMW) images and developed a concealed object detection system for AMMW image. WANG X L et al.[11] proposed patch based mixture of Gaussians that utilizes structure and uncertainty of objects to detect concealed items. TAPIA S L et al.[12] proposed a method that combines image processing and statistical machine learning techniques to solve localization/detection problem of passive millimeter-wave images.

Algorithms of object detection based on deep learning are mainly divided into two categories. One is a framework that combines object proposal, and the other is a framework for integrated convolutional networks. Faster R-CNN[13] is a detection method based on the regional proposal. Single Shot MultiBox Detector (SSD)[14] was proposed by WEI Liu et al. in 2016. It is a framework without a regional proposal. FPN[15] is proposed in a basic Faster R-CNN system for detecting objects at different scales.

However, the effect based on the traditional method mainly depends on the extracted features, but the low resolution and blurry characteristics of the millimeter-wave image have effect on feature extraction. Although feature extraction of convolutional neural networks is powerful, its robustness to distortion (rotation, scaling, etc.) is very low for small data sets in real-world detection. In the security check of the human body, objects carried by people will be presented in different sizes and angles. So far, there is no deep-learning-based method to study the distortion invariance in object detection of the millimeter-wave

image. In order to solve the problem of detecting objects of different sizes, shapes, and angles in a millimeter-wave image, a conventional method is generally employed. For example, low order Hu moments and other four shape features are combined together[16], choose saliency, SIFT and HOG features to form image descriptors[9], based on Haar-like features for PMMW image contraband detection[8]. Traditional methods based on feature descriptors are effective for some simple image classifications. But traditional methods are not up to the complicated situation. In addition, feature extraction by traditional methods is limited by low-quality millimeter-wave images.

Because of the these problems, we propose a detection framework of Spatial-Transformer Regional Convolutional Neural Network (ST-RCNN) with deblurring, combining non-blind deblurring and Spatial Transformer Networks (STN)[17].

1 Active millimeter-wave cylindrical scanning imaging

The antenna array rotates around the axis of the human body at the circumference of the radius ρ_a, and form a synthetic aperture in the direction of the circumference φ_a, $\varphi_a \in [0, 2\pi)$. Its position is $r_a = (x_a, y_a, z_a)^T$. The coordinate of the imaging point is defined as $r_o = (x_o, y_o, z_o)^T$, where $|r_o| \leqslant D$, D is the imaging area of the object. The length of the antenna array is L_{z_a}. The cylindrical aperture data during imaging is formed by sampling in directions of φ_a and z_a. The imaging frame is shown in Fig. 1.

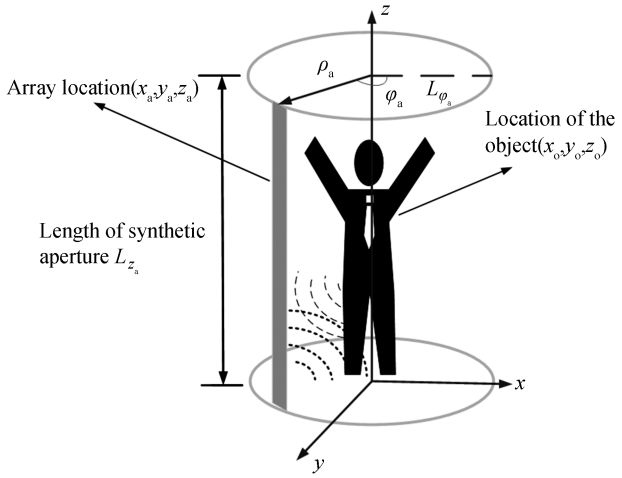


Fig.1 Model of millimeter-wave cylindrical scanning imaging

The Green's function of the spatial domain is defined as

$$G(\boldsymbol{r}, \boldsymbol{r}') = \frac{\mathrm{e}^{-\mathrm{j}k(r-r')}}{4\pi|\boldsymbol{r}-\boldsymbol{r}'|} = \frac{-1}{2\pi}\iiint_k \frac{\mathrm{e}^{-\mathrm{j}[k_x(x-x')+k_y(y-y')+k_z(z-z')]}}{k^2-k_x^2-k_y^2-k_z^2}\mathrm{d}k_x\mathrm{d}k_y\mathrm{d}k_z \tag{1}$$

where the wave number of the free space is represented by k_0, $k_0^2 = k_x^2 + k_y^2 + k_z^2$. $\boldsymbol{r}$ represents the position of the observation point, $\boldsymbol{r}'$ represents the object position. When the excitation field u_i and the scattering field u_s are known, $u_i \gg u_s$ and $o(\boldsymbol{r}) \ll 1$, the nonlinear inverse scattering equation can be linearized. The solution of linear inverse scattering is defined as

$$u_s(\boldsymbol{r}_a) = \iiint_{|r_o| \leqslant D} o(\boldsymbol{r}_o)\, u_i(\boldsymbol{r}_o)\, G(\boldsymbol{r}_a, \boldsymbol{r}_o)\, \mathrm{d}\boldsymbol{r}_o \tag{2}$$

where $\mathrm{d}\boldsymbol{r}_o = \mathrm{d}x_o\mathrm{d}y_o\mathrm{d}z_o$, It can be extended to monostatic radar, and its scattering field can be expressed as

$$u_s(\boldsymbol{r}_a) = \iiint_{|r_o| \leqslant D} o(\boldsymbol{r}_o)\, G(\boldsymbol{r}_a, \boldsymbol{r}_o)\, \mathrm{d}r_o \tag{3}$$

The scattering field u_s is obtained by linearly accumulating the weight of the objective function $o(\boldsymbol{r})$ and the Green's function $G(\boldsymbol{r}_a, \boldsymbol{r}_o)$ at the observation point $\boldsymbol{r}_a$. According to the data acquisition method of the millimeter wave imaging system, the polar cylindrical aperture coordinates are defined as (ρ_a, φ_a, z_a). Eq. (1) is substituted into Eq. (3) to obtain Eq. (4).

$$u_s(\varphi_a, z_a, \omega) = \frac{-1}{(2\pi)^3}\iiint_{|r_o| \leqslant D} o(\boldsymbol{r}_o)\iiint_k \frac{\mathrm{e}^{-\mathrm{j}[k_x(x_a-x_o)+k_y(y_a-y_o)+k_z(z_a-z_o)]}}{4k^2-k_x^2-k_y^2-k_z^2}\mathrm{d}k_x\mathrm{d}k_y\mathrm{d}k_z\mathrm{d}x_o\mathrm{d}y_o\mathrm{d}z_o \tag{4}$$

k_x, k_y, k_z replaced by the polar cylindrical coordinates (k_r, ϕ, k_z) in the spatial frequency domain, where

$k_x=k_r\cos\phi$, $k_y=k_r\sin\phi$, $k_r=\sqrt{k_x^2+k_y^2}$, $\phi=\arctan(k_y/k_x)$. Then Eq. (5) is obtained by ignoring all amplitude terms and passing the inverse Fourier transform on k_z and Fourier transform on ϕ

$$u_s(k_\phi,k_z,\omega)=\int o(k_r,k_\phi,k_z)\cdot F_\phi\{e^{-jk_r\rho_a\cos\phi}\}k_r dk_r \tag{5}$$

The algorithm of Stolt-mapping can be used for interpolation operations. Finally, the objective function is

$$o(x_o,y_o,z_o)=F^{-1}_{k_x,k_y,k_z}\left\{\text{Stolt}-\text{Mapping}\left[F^{-1}_{k_\phi}\left\{\frac{u_s(k_\phi,k_z,\omega)}{F_\phi\{e^{-jk_r\rho_a\cos\phi}\}}\right\}\right]\right\} \tag{6}$$

2 Structure of proposed network

Based on the principle of active millimeter-wave cylindrical scanning imaging, a millimeter-wave wideband T/R transceiver (frequency range from 24 GHz to 30 GHz), single-shot single-receiver wideband switch array, three-dimensional real-time imaging and detection framework of our ST-RCNN with deblurring are used to achieve fast 3D imaging of high resolution and object detection of dangerous goods. The object detection frame is more robust to translation, scaling, and rotation. Fig. 2 shows the detection framework of ST-RCNN with deblurring. It can be divided into two parts, the part of non-blind deblurring and the part of ST-RCNN detection.

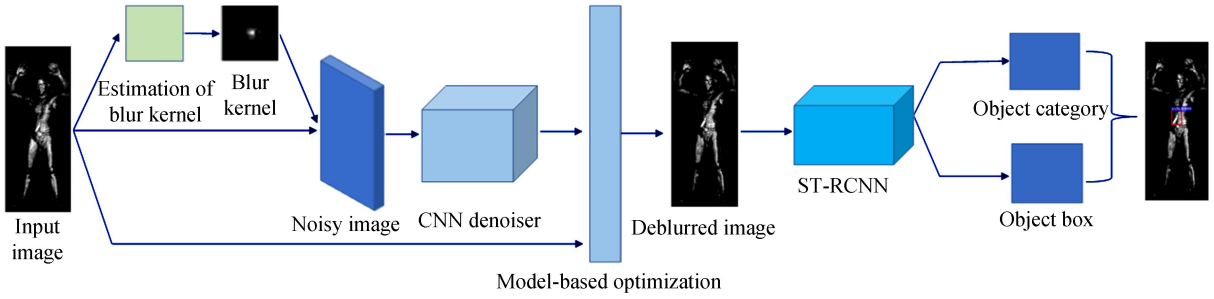


Fig.2 Frame for object detection of millimeter-wave images

2.1 Network of non-blind deblurring

At present, the quality of millimeter-wave imaging is generally low, with more noise and poor contrast. The deblurring of the millimeter-wave image should meet the requirements: the details of the image are enriched as much as possible, and the clarity should be enhanced; losing important features contained in the image should be avoided; no additional noise should be introduced during this process. To improve the quality of the millimeter-wave image, the image can be used for subsequent feature extraction and image recognition.

The overall non-blind deblurring architecture is shown in Fig. 3. First, the blur kernel of the original image is estimated. Then the original image is applied with a blur kernel and Gaussian noise is added to synthesize the blurred image. It is input to the CNN denoiser to get the image priori knowledge. Finally, the deblurred millimeter-wave image is obtained by the Half Quadratic Splitting (HQS) framework.

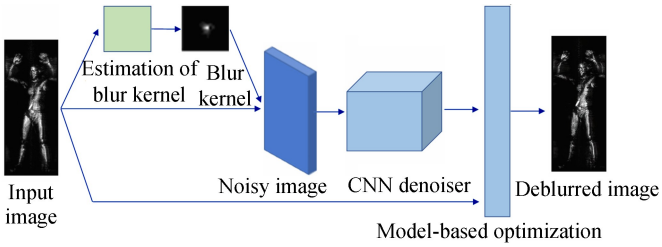


Fig.3 Network of non-blind deblurring

2.1.1 *Estimation of blur kernel*

The blurred image can be regarded as a convolution operation by the latent image and the blurring kernel, and the blur process is referred to

$$B=I\otimes K+N \tag{7}$$

where B is a blurred image, I is a latent image, K is a blur kernel, N is additional noise, and $\otimes$ is a

convolution operation.

The blur kernel is estimated as follows:

Step 1: Roughly initializing the value of K for the input blurred image.

Step 2: Solving Eq. (8) by minimizing I, u and g while determining other variables, and Eq.(9) by Fast Fourier Transform (FFT).

$$J(I,g,u)=\min_{I,g,u}\|I\otimes K-B\|_2^2+\alpha\|\nabla I-g\|_2^2+\beta\|D(I)-u\|_2^2+\lambda\|g\|_0+\omega\|u\|_0 \tag{8}$$

$$J(K)=\min_{K}\|\nabla I\otimes K-\nabla B\|_2^2+\gamma\|K\|_2^2 \tag{9}$$

where α and β are penalty parameters, u and g are auxiliary variables, λ, γ and ω are weight parameters, and $D(I)$ is a norm. Given I, the solution of u is

$$u=\begin{cases}D(I), & |D(I)|^2\geqslant\dfrac{\omega}{\beta}\\ 0, & \text{otherwise}\end{cases} \tag{10}$$

Eq. (10) is similar for the solution of g. The first term in Eq. (8) limits the convolution result of the latent image and the blur kernel to be similar to the input blur image. The second term retains a large gradient on the image gradient.

Step 3: Iterating Step 2 for i times, and finally the blur kernel K and the intermediate value of the latent image I are output.

2.1.2 *Network of convolutional denoiser*

A convolutional neural network combined with a model-based optimization is to achieve deblurring of millimeter-wave images. The noise image is synthesized by first applying a blur kernel and then adding additive Gaussian noise with a noise level σ. Then it is input to the CNN to get the priori information. It is integrated into a model-based optimization method to achieve deblurring of millimeter-wave images. The convolutional Denoiser is shown in Fig. 4.

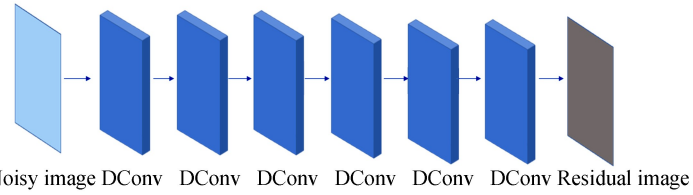


Fig.4 Structure of convolutional denoiser network

Among them, to capture content information, the receptive field is expanded by continuous dilated convolution. The CNN consists of 7 layers of dilated convolution and the dilation factors are set to 1, 2, 3, 4, 3, 2, 1, in turn. The dimension of the feature map is 64. The activation functions used are all ReLU functions. Batch normalization is used from the second layer.

The obtained image prior information is integrated into the model-based optimization algorithm to achieve image deblurring.

2.2 Detection framework of ST-RCNN

The STN is inserted into the feature extraction network VGG16[18] of the traditional Faster R-CNN[13], and a detection framework of ST-RCNN is obtained. It is shown in Fig. 5. The transformation mode of the space transformer depends on each sample, and each input sample can generate a suitable spatial transformation for its image or feature map, so that the Faster R-CNN is more robust to the translation, scaling and rotation of the sample data.

The basic network for feature extraction of Faster R-CNN is VGG16 with 13 convolutional layers. The first two layers separately contain 64 convolution filters followed by a max-pooling layer. Next are two convolutional layers containing 128 convolution filters and a max-pooling layer. Three convolutional layers and a pooling layer are performed. The last six convolutional layers contain 512 convolutional filters respectively and are connected to the max-pooling layer after the tenth and thirteenth convolutional layers. The obtained feature map is input into the Region Proposal Networks(RPN) to obtain Region of Interest (RoI). And then combined with the feature map, the result is output after the RoI pooling layer and the Fully-Connected (FC) layers. The STN network is combined to the first layer of the convolutional layer of VGG16. The RELU function is used as a nonlinear activation function.

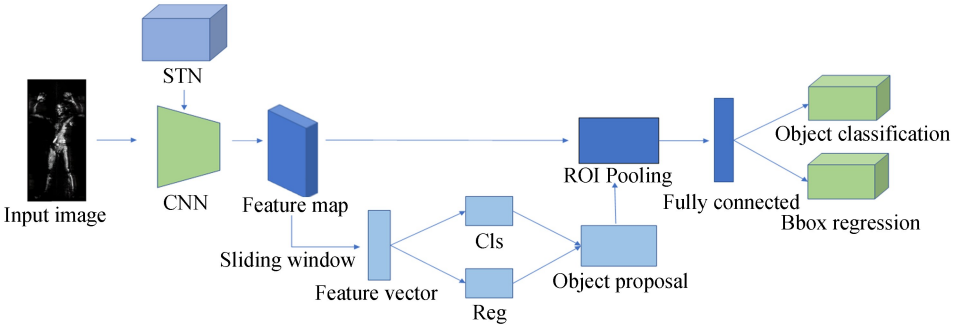


Fig.5 Structure of ST-RCNN

2.2.1 *Spatial-transformer network*

The spatial transformer network is mainly composed of three parts: parameter prediction, coordinate mapping, and pixel acquisition. The main idea is to adjust the weight and perform an affine transformation to achieve the purpose of scaling and rotation, as shown in Fig. 6. The module of spatial transformation can be inserted into the convolutional neural network to automatically learn how to implement the transformation of the feature map. Eventually, the network model learns the invariance of translation, scaling translation, rotation, and more common distortions.

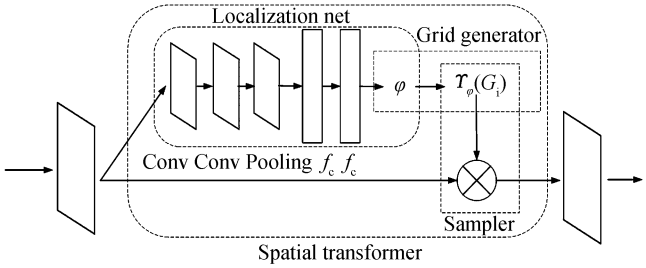


Fig.6 Structure of spatial transformer networks

The parameters φ of transformation obtained by localization net. The structure of feature extraction contains two convolutional layers, followed by the max-pooling layer. Then transform parameters are obtained by regression of two-layer FC layer. The activation function is the ReLU function. The Grid generator solves the feature map between the output and the input by the parameter φ and the transformation mode. The sampler combines the feature coordinates and parameters φ to select the input features and combines the bilinear interpolation to obtain the results.

2.2.2 *Localization network*

The prediction of parameters is implemented through the localization network in the STN. Rotation is a kind of affine transformation. As shown in Fig. 7, point $A(x_a, y_a)$ is rotated by a θ degree angle to obtain point $B(x_b, y_b)$. The corresponding relationship is

$$x_a = x_b\cos\theta - y_b\sin\theta \tag{11}$$

$$y_a = y_b\cos\theta + x_b\sin\theta \tag{12}$$

where $\cos\theta$ and $\sin\theta$ are weight parameters. They are performed in a matrix form to complete the rotation.

$$\begin{bmatrix} x_b \\ y_b \end{bmatrix} = \begin{bmatrix} \varphi_{11} & \varphi_{12} \\ \varphi_{21} & \varphi_{22} \end{bmatrix} \begin{bmatrix} x_a \\ y_a \end{bmatrix} + \begin{bmatrix} \varphi_{13} \\ \varphi_{23} \end{bmatrix} \tag{13}$$

where $\varphi_{11} = \cos\theta, \varphi_{12} = -\sin\theta, \varphi_{21} = \sin\theta, \varphi_{22} = \cos\theta, \varphi_{13} = \varphi_{23} = 0$. And φ_{13}, φ_{23} are offset parameters.

(a) Point A is rotated to obtain Point B (b) Rotation of the object

Fig.7 Transformation of rotation

For scaling, it is performed in matrix form as

$$\begin{bmatrix} x_b \\ y_b \end{bmatrix} = \begin{bmatrix} \varphi_{11} & 0 \\ 0 & \varphi_{22} \end{bmatrix} \begin{bmatrix} x_a \\ y_a \end{bmatrix} + \begin{bmatrix} \varphi_{13} \\ \varphi_{23} \end{bmatrix} \tag{14}$$

where $\varphi_{12}=\varphi_{21}=0$, and φ_{13}, φ_{23} are offset parameters.

The feature map can be taken as input by the localization network, and then the parameters are returned through convolution, full connection layer, and so on. By determining the 6-dimensional parameters of affine transformation, these operations of affine transformation can be implemented by the following steps.

2.2.3 *Grid generator*

According to the principle of affine transformation, the feature map is transformed by the grid generator using the predicted parameters. The relationship of coordinate matrix transformation between the target image and the original image is defined as

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \boldsymbol{Y}_\varphi(G_i) = \begin{bmatrix} \varphi_{11} & \varphi_{12} & \varphi_{13} \\ \varphi_{21} & \varphi_{22} & \varphi_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \tag{15}$$

where the coordinate point of the original image is represented as (x_i^s, y_i^s), the output image coordinate point is represented as (x_i^t, y_i^t), and $\boldsymbol{Y}_\varphi(G_i)$ represents the mapping matrix containing the relationship of an affine transformation. By learning the six parameters φ_{11}, φ_{12}, φ_{13}, φ_{21}, φ_{22}, φ_{23} of the affine transformation matrix, the invariance of transformations such as rotation, scaling, and translation can be improved by the CNN.

2.2.4 *Sampler*

After passing through the localization network and the grid generator, the output feature map is mapped to the input feature map by the spatial transformation. The parameters cannot be trained by the STN through back-propagation when only the above two processes are included. The backpropagation condition is satisfied by the bilinear interpolation of the sampling network. Its formula for bilinear interpolation is

$$V_i^C = \sum_n^H \sum_m^W U_{nm}^c \max(0, 1 - |x_i^s - m|) \max(0, 1 - |y_i^s - n|) \tag{16}$$

where V_i^C represents the output value for pixel i at location (x_i^t, y_i^t) in channel C. U_{nm}^C represents the score value of the coordinate (n, m) in the color channel C, and the max function can be guided. The parameters can be updated by Eq.(16) through backpropagation.

3 Experimental results

Detection framework of our ST-RCNN with deblurring can achieve high resolution imaging and object detection of dangerous goods.

The millimeter-wave imaging system based on cylindrical rotation scanning is mainly for human body security, with an imaging distance of 68 cm and an imaging area of 100 cm × 210 cm. The three-dimensional millimeter-wave data of the human body is collected by the system, and the three-dimensional circular scan image of the human body is obtained by data signal processing technology. 200 images were randomly picked from the images obtained by multiple imaging for testing. The image is then deblurred and object detected.

A millimeter-wave image dataset for human body security of object detection have been established. The image is labeled as the VOC2007 standard and the objects are marked with a knife and a gun. Hardware configuration of experimental server is 2×CPU(E5-2609 v4), 64 GB memory, Tesla K80.

3.1 Result of deblurring networks

Millimeter-wave images are processed by five deblurring methods. Blurred images captured by millimeter-wave devices are especially challenging for most deblurring methods. Deblurred results are obtained through advanced methods [4], [19-21]. As shown in Fig. 8, the image restored by our method is superior to the recent deblurring methods.

The results of the Peak Signal to Noise Ratio (PSNR) for millimeter-wave images are summarized. The value of PSNR is usually referred to measure the quality of the processed image. If the value of PSNR is larger, it means less distortion. Its formula is shown as

$$\mathrm{PSNR} = 10 \times \log_{10}\left[\frac{(2^n-1)^2}{\mathrm{MSE}}\right] \tag{17}$$

where MSE is the mean square error between the original image and the processed image. The average PSNR of our method is higher than other methods of deblurring. It is at least 1.64 dB higher than the other four deblurring methods, as shown in Table 1.

(a) Blurred image (b) Result of Ref. [19] (c) Result of Ref. [20] (d) Result of Ref. [4] (e) Result of Ref. [21] (f) Our result

Fig.8 Millimeter-wave image and its corresponding result of deblurring

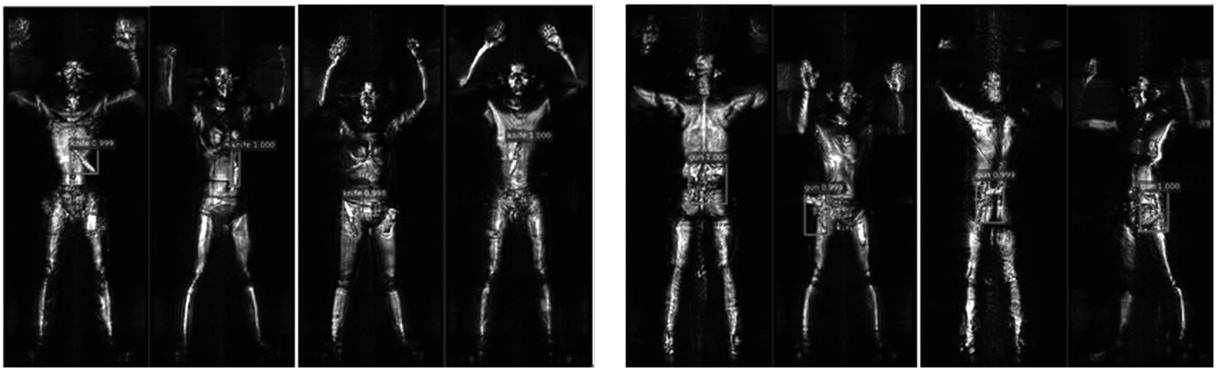
Table 1 Quantitative evaluation of deblurring on millimeter-wave images

Method	Ref.[19]	Ref.[20]	Ref.[4]	Ref.[21]	Ours
PSNR/dB	22.80	23.41	23.82	25.85	27.49

3.2 Comparison with state-of-the-art methods

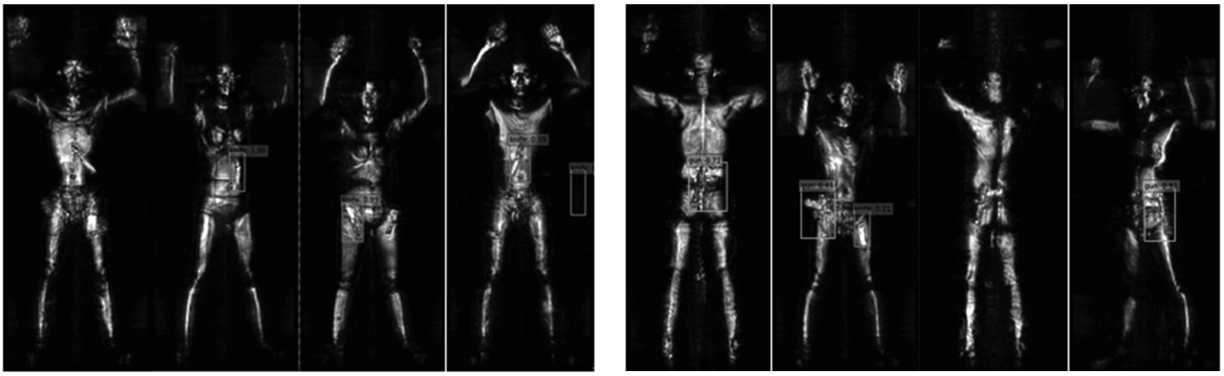
The feature extraction networks of Faster R-CNN (VGG)[13], SSD[14], FPN[15] and ST-RCNN are based on VGG[18] structure D (VGG16), which includes a 5-segment convolutional layer and the FC layers. It consists of 13 convolutional layers with 3×3 convolution filters, 5 max-pooling layers with 2×2 convolution filters, and 3 FC layers. Faster R-CNN performs feature extraction through the basic layers, including 13 convolutional layers and ReLU layers, and 4 max-pooling layers. The resulting feature map is used for subsequent RPN and RoI pooling layer. After the RoI Pooling layer, a fixed-size proposal feature map is obtained, which is identified and located after the FC layer. In SSD, the FC layer of VGG16 is replaced by 2 convolutional layers, adding 4 convolutional layers. Feature maps of different scales are extracted from the 4th and 7th to 11th convolutional layers, and finally the feature maps of different scales are predicted separately. FPN is based on Faster RCNN. In the forward process, the feature maps obtained by different convolutional layers form a feature pyramid. The feature map of the last convolutional layer is upsampled by interpolation. The feature map of the feature pyramid is convolved with the 1×1 convolution filters, and is merged with the corresponding upsampled feature map, and the fusion result is convoluted by the 3×3 convolution filters to obtain a new feature map. The feature extraction network of Faster R-CNN (ResNet) is ResNet101[22]. It consists of a $7\times7\times64$ convolutional layer, 33 building blocks with 3 convolutional layers, and a FC layer.

The result images of the object detection are obtained by the SSD method and the proposed method. Compared with the SSD method , better results can be obtained by the proposed method . Fig . 9 shows the



(a) Object detection result of the knife (b) Object detection result of the gun

Fig.9 Example of results in a millimeter-wave image dataset using our method



(a) Object detection result of the knife (b) Object detection result of the gun

Fig.10 Example of results in a millimeter-wave image dataset using SSD network

results of object detection in a millimeter-wave image dataset using our method. Knives and guns of different angles and sizes can be accurately detected. Fig. 10 shows the results of the object detection obtained by the SSD method. It contains some images of wrong and missing detection. The knife carried by the first person in Fig. 10 (a) and the gun carried by the third person in Fig. 10 (b) were missed. The objects of the fourth person in Fig. 10 (a) and the second person in Fig. 10 (b) are detected by mistake. However, the proposed method can accurately detect the images of wrong and missing detection.

The value of Average-Precision (AP) can be calculated from the area under the precision-recall curve. Mean Average Precision (mAP) is the average of AP values for multiple categories. As can be seen from Table 2, the proposed method is 8.9% and 4.6% more accurate than SSD and FPN, respectively. Compared with the Faster R-CNN (VGG), the proposed method is improved by 4.4% mAP. For knife and gun detection, the proposed method is 7.6% and 1.3% better in AP, respectively Compared to Faster R-CNN (ResNet), the mAP of the proposed method increases by 19%.

Table 2 Comparison of object detection performance in different network

	AP for knife/%	AP for gun/%	mAP/%
Faster R-CNN[13] (VGG)	73.0	80.0	76.5
Faster R-CNN[13] (ResNet)	52.9	70.8	61.9
SSD[14]	71.4	72.7	72.0
FPN[15]	74.9	77.8	76.3
ST-RCNN	78.4	81.0	79.7
Proposed method	80.6	81.3	80.9

3.2.1 *Real-time analysis of algorithms*

Faster R-CNN is a detection algorithm based on object proposal. It can achieve near real-time detection speed. SSD is a detection algorithm based on an integrated convolutional network. The result can be obtained directly after a single detection, so its speed is faster than that of Faster R-CNN, but the accuracy of detection is lower. FPN is an algorithm of multi-scale object detection whose prediction is performed independently at different feature layers, and its detection speed is slower than that of Faster R-CNN.

For 2D millimeter-wave images with a size of 205×512, the running time of the non-blind deblurring algorithm is about 10 ms in the case of GPU acceleration. Non-blind deblurring is an algorithm of image preprocessing. For the speed of detection, it takes about 50 ms to run the ST-RCNN model, which is 6% slower than the Faster R-CNN. For the same image, preprocessing and detection are performed serially, with the total detection time of about 60 ms. In channel-based security, pedestrians spend approximately 2 s in a rotating scan, performing active 3D imaging and projecting 2D images of different angles at approximately 10 frames/s. Image processing should take less than 100 ms. The total time for image processing is about 60 ms, which is less than 100 ms, so our solution can meet the real-time requirements of channel security. In practical applications, pre-processing and network detection can be performed in two phases, which can improve system throughput.

3.2.2 *The effect of STN on detection accuracy*

The millimeter-wave image without preprocessing is directly detected by the Faster R-CNN, with a

mAP of 76.5%. The millimeter-wave image without preprocessing is directly detected by the network of ST-RCNN, with a mAP of 79.7%. For the object detection network of millimeter-wave images, its mAP can be improved by 3.2% through STN.

3.2.3 *The effect of deblurring on detection accuracy*

The millimeter-wave image is directly detected by ST-RCNN with the mAP of 79.7%. The millimeter-wave image is directly detected by the network of ST-RCNN with deblurring, and the mAP is 80.9%. The mAP can be improved by 2.2% to 80.9% through the network of non-blind deblurring.

4 Conclusion

Due to the low resolution of millimeter-wave images, and the spatial transformation of the object, such as rotation, scaling, and translation, the object detection is more challenging.For the small data sets, according to the characteristics of millimeter-wave images, a detection framework of spatial-transformer RCNN with deblurring is proposed, which combines non-blind deblurring and spatial transformer networks. This paper implements the object detection of millimeter-wave image hiding dangerous items. Secondly, the image is enhanced to enrich the details of the image. Compared with some state-of-the-art methods, image quality and detection accuracy can effectively be improved. We hope to introduce multi-scale features into our method to solve the problem of detection for the smaller objects in the millimeter-wave images in the continuing work.

References

[1] YI D, KIM S, YEOM S, *et al*. Experimental study on image interpolation for concealed object detection[C]. MFI, Daegu: IEEE, 2017: 501-504.

[2] MATEOS J, LÓPEZ A, VEGA M, *et al*. Multiframe blind deconvolution of passive millimeter wave images using variational dirichlet blur kernel estimation[C]. ICIP, Arizona: IEEE, 2016: 2678-2682.

[3] RAJAGOPALAN A N. Motion deblurring: algorithms and systems[M]. MADRAS, CHELLAPPA R. Cambridge: Cambridge University Press, 2014.

[4] PAN Jin-shan, HU Zhe, SU Zhi-xun, *et al*. Deblurring text images via L0-regularized intensity and gradient prior[C]. CVPR, Columbus: IEEE, 2014: 2901-2908.

[5] PAN Jin-shan, HU Zhe, Su Zhi-xun, *et al*. Deblurring face images with exemplars[C]. ECCV, Zurich: Springer International Publishing AG, 2014: 47-62.

[6] HU Zhe, CHO Sunghyun, WANG Jue, *et al*. Deblurring low-light images with light streaks[C]. CVPR, Columbus: IEEE, 2014: 3382 - 3389.

[7] VASU S, MALIGIREDDY V R, RAJAGOPALAN A N. Non-blind deblurring: handling kernel uncertainty with CNNs [C]. CVPR, Salt Lake City: IEEE, 2018: 3272-3281.

[8] XIAO Ze-long, LU Xuan, YAN Jiang-jiang, *et al*. Automatic detection of concealed pistols using passive millimeter wave imaging[C]. IST, Macau:IEEE , 2015: 1-4.

[9] LI Zheng, JIN Ying-kang, SHEN Zong-jun, *et al*. A synthetic targets detection method for human millimeter-wave holographic imaging system[C]. CCBD, Macau: IEEE, 2015: 284-288.

[10] LIU Ting, ZHAO Yao, WEI Yun-chao, *et al*. Concealed object detection for activate millimeter wave image[J]. *IEEE Transactions on Industrial Electronics*, 2019, **66**(12): 9909-9917.

[11] WANG Xin-lin, GOU Shui-ping, WANG Xiu-xiu, *et al*. Patch-based gaussian mixture model for concealed object detection in millimeter-wave images[C]. TENCON 2018 - 2018 IEEE Region 10 Conference, Jeju: IEEE ,2018: 2522-2527.

[12] TAPIA S L, MOLINA R, BLANCA N P d l. Detection and localization of objects in passive millimeter wave images [C]. EUSIPCO, Budapest: IEEE 2016: 2101-2105.

[13] REN Shao-qing, HE Kai-ming, GIRSHICK R. Faster R-CNN: towards real-time object detection with region proposal networks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(6): 1137-1149.

[14] LIU Wei, ANGUELOV D, ERHAN D, *et al*. SSD: single shot multibox detector[C]. ECCV, Amsterdam: Lecture Notes in Computer Science, 2016: 21-37.

[15] LIN Tsung-yi, DOLLÁR P, GIRSHICK R, *et al*. Feature pyramid networks for object detection[C]. CVPR, Honolulu: IEEE, 2017: 936-944.

[16] DAI Ling, HU Hong, CHEN Yi-fan, *et al*. Millimeter-wave image target recognition based on the combination of shape features[C]. ICIA, Ningbo: IEEE 2016: 1732 - 1736.

[17] JADERBERG M, SIMONYAN K, ZISSERMAN A, *et al*. Spatial transformer networks[C]. Montreal: NIPS, 2015: 2017-2025.

[18] SIMONYAN K , ZISSERMAN A . Very deep convolutional networks for large-scale image recognition[J]. *Computer Science*, 2014, arXiv:1409.1556v6.

[19] KRISHNAN D, TAY T, FERGUS R. Blind deconvolution using a normalized sparsity measure[C]. CVPR, Colorado Springs: IEEE, 2011: 233-240.

[20] PAN Jin-shan, LIU Ri-sheng, SU Zhi-xun, *et al*. Motion blur kernel estimation via salient edges and low rank prior [C]. ICME, Chengdu: IEEE ,2014: 1-6.

[21] PAN Jin-shan, SUN De-qing, PFISTER H, *et al*. Blind image deblurring using dark channel prior[C]. CVPR, Las Vegas:IEEE, 2016: 1628-1636.

[22] HE Kai-ming, ZHANG Xiang-yu, REN Shao-qing, *et al*. Deep residual learning for image recognition[C]. CVPR, Las Vegas:IEEE, 2016: 770-778.