A Survey of Surface Defect Detection Based on Deep Learning

Weihua Yang(✉)

Tianjin University of Technology and Education, Tianjin, People's Republic of China
xiaohua007qaq@outlook.com

Abstract. In recent years, with the rapid development of technologies such as computers and artificial intelligence, various research fields based on deep learning have been broadly used, among which industrial detection is the most important. In this paper, the definition of defects and defect detection is firstly defined. Then, several mainstream methods of surface defect detection based on convolutional neural network are introduced in recent years, and the typical application scenarios of each method are summarized. Finally, two key problems in surface defect detection are discussed.

Keywords: Machine vision · Surface defect detection · Deep learning · Convolutional neural network (CNN)

1 Introduction

Surface defect detection may be a vital analysis content within the sphere of machine vision, additionally referred to as AOI (Automated Optical Inspection) or ASI (Automated Surface Inspection). It is the use of machine vision equipment whether there are any defects in judging image collection and image technology. Nowadays the surface defects of equipment based on machine vision has been widely utilized in various industrial fields instead of human eye detection, mainly including automobile, machinery manufacturing, household appliances, semiconductor and electronics, aerospace, chemical and numerous others. The conventional defect detection methodology supported machine vision typically adopts standard image process algorithmic rule or artificial style feature and classifier. In the design of traditional detection methods, the use of carefully constructed image processing methods can greatly reduce the design difficulty of traditional methods and increase usage costs. However, in practical and complex practical applications, the identification of surface defects often encounters problems such as small difference between image and background, low contrast, and large defect scale variation. In this case, the classical methods are often helpless and difficult to achieve good detection results.

Recently, deep learning algorithm models based on convolutional neural network (CNN) have been widely utilized in numerous computer vision fields. For example, face recognition, target tracking and automatic driving, etc. In different industrial environments, defect recognition technology based on deep learning has also been broadly

C. F. Peng et al. (Eds.): MMET 2022, ASSEHR 703, pp. 362–367, 2023.
https://doi.org/10.2991/978-2-494069-51-0_51

used. As a result, the surface defect detection approach based on deep learning offers significant academic research value as well as a very wide business application promise.

The purpose of this paper is to provide researchers with a better comprehension of the commonly used surface defect detection techniques through a review of domestic and foreign literature on surface defects. The substance of this paper is orchestrated as takes after: First, the definition of defect detection problem is explained. Then, the characteristics of convolutional neural network (CNN), autoencoder neural network, residual convolution neural network (ResNet) and full convolution neural network (FCN) are introduced, and their applications in surface defect detection are introduced. Finally, the two key problems of small sample size and real-time in surface defect detection are discussed.

2 Defect Detection Problem

2.1 Definition of Defects

In machine vision work, defects are often determined by human experience rather than purely mathematical definitions. Two entirely distinct detection techniques will result from different cognitions of the defect mode. The first is a supervised way, which is trained by feeding the network with labeled defective pictures. At this time, defects refer to marked areas or images. Thus, more research has been done on the characteristics of defects. The second category is unsupervised defect identification. In general, network training requires normal, error-free samples. This method focuses more on flawless properties. When an unseen feature is found in the process of defect detection, a defect is considered to be detected. In this case, defect means exception, so this approach is also called Anomaly detection.

2.2 Definition of Defect Detection

In contrast to the clear classification, detection and segmentation tasks in computer vision, the necessities of defect detection are a little vague. In fact, we will divide requirements into three steps. The first step, corresponding to the classification work in computer vision, the work in this stage is called "defect classification", which only provides the classification info of the image. The second step, which corresponds to the localization work in computer vision. This method can not only obtain the defect type in the image, but also locate the defect accurately. The third step, resembling the segmentation work in computer vision, which separates the defect from the background, so as to get a number of data such as the location, length and area of this defect. Although the functional requirements and purposes of defect detection in the three steps are different, they actually contain each other and can be converted into each other. Therefore, in accordance with the traditional industry practice, it is collectively called defect detection, which just depends on the network, and the different functions.

3 Defect Detection Technology

3.1 Convolutional Neural Network

CNN, a feedforward neural network, is one of the most popular Neural Network structures in deep learning. Common CNN network structures include multiple convolutional layers and pooling layers, which are used for image feature extraction. Feature images are nonlinear processed by activation functions and feature maps are transmitted to the next layer. Then feature classification is carried out by full connection layer and Softmax structure. CNN includes three features of weight sharing in structure, local connection and spatial or temporal down-sampling, which effectively reduce the complexity of network model parameters.

There are also many cases of convolutional neural networks applied to surface defect detection. Weimer et al. [1] replaced the traditional industrial vision detection method with deep convolutional neural network to overcome the traditional method's high dependence on manual experience to select defect features, and verified that the convolutional neural network method achieved better results. Yao et al. [2] proposed a new CNN structure for the study of track defect detection, and experiments proved that the network structure has a high recognition rate in track surface defect recognition.

3.2 Self-coding Network

Encoding and decoding are the two fundamental processes of autoencoder neural network. The input signal is transformed into an encoding signal during the encoding stage in order to extract features. Feature information is transformed into reconstructed signals during the decoding stage, and defect detection is made possible by minimizing the reconstruction error through weight and bias adjustments. The autoencoder neural network's learning objective is feature learning rather than classification, which is the main distinction between it and other machine learning techniques. It is very capable of nonlinear mapping and autonomous learning, and it can learn nonlinear metric functions to tackle complex background issues.

Tao et al. [3] proposed CASAE algorithm. In this algorithm, an autoencoder is used to design a cascade self-encoder network structure based on its good performance in image reconstruction. Yun et al. [4] proposed a new conditional convolutional variational autoencoder (CCVAE) and deep convolutional network (DCNN) classification algorithm for metal surface defect detection.

3.3 Residual Convolution Neural Network

As the network depth increases, the features of CNN and generative adjunctive network [5] increase, but at this time, it is so easy to cause non-convergence of activation function. The goal of this method is to use the residual optimization method to increase the quantity of layers of the network on the basis of maintaining the quantity of network layers, so that the size of the convolutional layer and therefore the input unit in the remaining units are a similar, and thru the activation perform to cut back losses.

Chen et al. [6] designed a fabric defect detection system using the Faster R-CNN architecture. In order to obtain the ability of shallow network to extract small-scale defects, residual network got utilized to improve the Faster R-CNN feature extraction network.

3.4 Full Convolution Neural Network

The usual CNN network will add multiple full connected layers after the convolution layer, so that the feature map generated by the convolution layer can be converted into a feature vector with constant length. However, the FCN network can settle for input images of any size, and also the deconvolution layer is used to up-sample the feature map of the last convolutional layer to make it precisely the same size as this input image. Therefore, each image all elements have a predictive effect. Pixel classification of up-sampling feature maps while conserving the spatial information of the input image.

He et al. [7] adopts Mix-FCN network. Behind the input layer is a series of convolution blocks (a, b, c, d, e, f and g) and up-sampling blocks (h). The first four convolution modules are maximized to reduce model parameters. The f convolution block uses two Inception convolution modules to capture multi-scale semantic information. The experimental results indicate that this new method can effectively improve the detection accuracy.

4 Key Problems

4.1 Small Sample Size Problem

At display, deep learning strategies are broadly utilized in different computer vision assignments, and surface imperfection discovery is for the most part respected as its particular application within the mechanical field. Since the number of industrial defect samples provided by deep learning algorithms in actual production is too small, it is impossible to directly detect surface defects. Compared with more than 140,000 samples on ImageNet, the main problem in the identification of surface defects is the small number of samples. In many practical industrial environments, the number of defective images is only a few or even dozens. In fact, there are four different solutions to the problem of small samples:

Data Amplification, Synthesis and Generation
The most common method of defect image amplification is to obtain more samples from original defect samples by image manipulations such as mirror, rotation, distortion, filtering, translation and contrast adjustment. For example, Wei et al. [8] and Li et al. [9] used the above method to apply amplified defect data to defect detection of textile surface and rail surface.

Network Pre-training or Transfer Learning
Due to the large number of parameters of depth learning network, it is simple to overfit by directly using small sample training network. But in the pre-training model, there

are certain common feature data and weight information. In 2018, Ren et al. [10] first performed transfer learning on surface defects, and their pre-training model adopted ImageNet pre-training model. Sun et al. [11] applied the transfer learning method to the surface defect detection of metal parts, which improved the accuracy and operation efficiency of model classification.

Design a Reasonable Network Structure
Reasonable design of network structure can also greatly reduce the need for samples. For example, Tabernik et al. [12] designed a multi-task defect detection network integrating classification and branch segmentation. Backbone, which shares feature extraction with two branches. This method can use each pixel in the image as a training sample for training the network. This greatly reduces the sampling requirements for the network.

Unsupervised and Semi-supervised Models
Both methods can reduce the need for samples. In the unsupervised case, the method only trains ordinary samples, so no erroneous samples are required. Under the condition of small samples, semi-supervised algorithms can use unlabeled samples for network training.

4.2 Real-Time Problem

In practical engineering, the defect identification technology based on deep learning is compartmentalized into three parts: data annotation, model training and model reasoning. In the real-time environment, people pay more and more attention to model reasoning. Currently, most defect detection methods mainly emphasize on the accuracy of defect classification and discrimination, while the inference efficiency of the model is very low.

At present, there are many ways to speed up models, such as model weight quantization and model pruning. Pan et al. [13] applied the FPGA accelerated Fourier reconstruction operator to texture surface defect segmentation, and the parallel acceleration structure of FPGA was three times that of the CPU of the same level. Although existing deep learning models use the GPU as a general-purpose computing unit, FPGA will become an attractive alternative as the technology evolves.

5 Conclusions

With the continuous development of modern artificial intelligence technology, surface defect recognition technology based on machine vision has also transferred from traditional machine learning, image processing and other fields to the field of deep learning. In this paper, the current commonly utilized deep learning algorithms are systematically summarized and analyzed, and the key of the question defect detection are discussed in depth, which provides a powerful reference for future research work.

References

1. Weimer, D., Scholz-Reiter, B., & Shpitalni, M. (2016). Design of deep convolutional neural network architectures for automated feature extraction in industrial inspection. CIRP Annals, 65(1), 417-420.
2. Yao, Z. W., Yang, H. F., HU, J. Y., Huang, Q. P., Wang, Z., & Bi, Q. S. (2021). Track Surface Defect Detection Method Based on Machine Vision and Convolutional Neural Network. Journal of the China Railway Society, 43(04), 101-107.
3. Tao, X., Zhang, D., Ma, W., Liu, X., & Xu, D. (2018). Automatic metallic surface defect detection and recognition with convolutional neural networks. Applied Sciences, 8(9), 1575.
4. Yun, J. P., Shin, W. C., Koo, G., Kim, M. S., Lee, C., & Lee, S. J. (2020). Automated defect inspection system for metal surfaces based on deep learning and data augmentation. Journal of Manufacturing Systems, 55, 317-324.
5. Lin, Y. L., Dai, X. Y., Li, L., Wang, X., & Wang, F. Y. (2018). The New Frontier of AI Research: Generative Adversarial Networks. Acta automatica Sinica, 44(05), 775-792.
6. Chen, K., Zhu, W., Ren, Z. F., & Zheng, Y. Y. (2020). Fabric Defect Detection Method Based on Deep Residual Network. Journal of Chinese Computer Systems, 41(04), 800-806.
7. He, T., Liu, Y., Xu, C., Zhou, X., Hu, Z., & Fan, J. (2019). A fully convolutional neural network for wood defect location and identification. IEEE Access, 7, 123453-123462.
8. Wei, B., Hao, K., Tang, X.-s., & Ding, Y. (2019). A new method using the convolutional neural network with compressive sensing for fabric defect classification based on small sample sizes. Textile Research Journal, 89(17), 3539.
9. Li, L. Y., Liu, Q., Zou, Y. M., Chen, J. Y., Li, P., & Wang, Q. X. (2021). Rail Surface Defect Detection Based on Improved YOLOv5 Algorithm. Journal of Wuyi University(Natural Science Edition), 35(03), 43-48+54.
10. Ren, R., Hung, T., & Tan, K. C. (2017). A generic deep-learning-based approach for automated surface inspection. *IEEE transactions on cybernetics, 48*(3), 929-940.
11. Sun, N., Xu, G. A., Wang, D. M., & Chen, T. (2022). Spot defect detection on metal surface based on transfer learning. Journal of Shanghai Dianji University, 25(02), 82-87+94.
12. Tabernik, D., Šela, S., Skvarč, J., & Skočaj, D. (2020). Segmentation-based deep-learning approach for surface-defect detection. Journal of Intelligent Manufacturing, 31(3), 759-776.
13. Pan, Y., Lu, R., & Zhang, T. (2020). FPGA-accelerated textured surface defect segmentation based on complete period Fourier reconstruction. Journal of Real-Time Image Processing, 17(5), 1659-1673.