Reading Aviation Wire Text in Natural Images under Assembly Workshop via Deeplearning

To cite this article: Shufei Li et al 2019 IOP Conf. Ser.: Mater. Sci. Eng. 563 042075

View the article online for updates and enhancements.
Reading Aviation Wire Text in Natural Images under Assembly Workshop via Deeplearning

Shufei Li, Lianyu Zheng*, Yiwei Wang and Renjie Zhang
School of Mechanical Engineering and Automation, Beihang University,
Beijing 100191, China
*Email: lyzheng@buaa.edu.cn

Abstract. Reading aviation wire text is an important step for aircraft wire assembly and also a challenging problem. Aviation wire text reading means to detect and recognize the serial number printed on the surface of aviation wires or aircraft cables. The main challenges of this task lie on small sizes, low contrast and background noise in images. In this paper, we propose a new method for aviation wire text detection and recognition via deep neural networks. The detection network utilizes text border features to roughly locate text regions in images. Then the coordinates of bounding boxes of texts are regressed by pixel-level predicted offsets in adjacent regions. The recognition network is capable to recognize the text regions with diverse lengths after padding strategies. In addition, a dataset of aviation wires captured by machine vision camera in the assembly workshop is developed. A flexible method to generate synthetic images of text whose characters are arranged in styles of both head-to-tail and left-to-right is presented. On aviation wire datasets, the proposed method for text detection and recognition achieves comparable performance and significantly outperforms previous approaches.

1. Introduction

The cost and workload for assembling aviation wires accounts for a large proportion in the process of aircraft assembly. Workers need to identify the serial numbers printed on the aviation wires and query the corresponding installation information from the database. It is difficult to adapt to the mass production of aircraft due to the ineffective productivity of manual identification. Technological innovations of digital image processing, pattern recognition, artificial intelligence etc. enable computer vision to read the aviation wire text automatically, and consequently, significantly improve the assembly efficiency of the aircraft. The traditional technology of optical character recognition (OCR) is used to solve the text recognition problem from the scans of document images, which have clean background. However, the recognition accuracy drops sharply when apply OCR method directly on the images of aviation wires that inevitably contain the environmental backgrounds of the manufacturing workshop.

The modern approach of deep learning comprising of two stages, i.e., text detection and text recognition, is employed to accurately identify the aviation wire text from images with the background of the manufacture workshop. Labels of each image are annotated to develop the original dataset. The detection model is trained on the prepared datasets to localize the text regions in images. Finally, the region of aviation wire text is identified by the recognition model. The remainder of the paper is organized as follows. Section II briefly reviews related work. Section III introduces details of our proposed method. The result on datasets and the rationality analysis of the performance is detailed in Section IV and conclusions are drawn in Section V.
2. Related work

Generally, text reading can be divided into two sub-stages: text detection and text recognition. The former aims to predict bounding boxes of the text from input images while the latter focuses on transforming the cropped text images into machine-interpretable character sequences. We cover both tasks in this paper while focus more on detection.

2.1. Text detection

The approach of scene text detection (referred to as EAST) [1] gives pixel-level text prediction in images. The work presents adaptation vertex-regressors for datasets in different annotation forms including rotating rectangular and quadrilateral labels of text regions in images. To address the problem of poor performance for long text detection in images, Chuhui Xue et al. [2] uses image augmented technology to extend the number of images in dataset. Specifically, they repeatedly sample text segments in different widths from the original images of dataset. The trained model fed by these augmented images shows outstanding performance in long text detection.

Inspired by the above method, we implement data augmentation to produce a large number of images having similar features to the original images of aviation wires. All the images are fed into detection network to improve the feature extraction ability of the model. Finally, the detection model outputs pixel-level text predictions, including the text score map, the text border map and the vertex offset. The position of the aviation wire text in images is regressed according to the output.

2.2. Text recognition

Jaderberg et al. verify that model trained with synthetic images can be successfully applied to the text identification problems of real-world scenes [3]. The text recognition network (referred to as CRNN) [4] is capable of identifying multi-length text sequences and is trained end-to-end. The model contains layers of bidirectional long short term memory (BLSTM) and connectionist temporal classification (CTC) [9] to learn the contextual relationship of character sequences in images. Thereby the CRNN achieves a very competitive accuracy for text recognition compared with approach in the past.

Considering that the amount of image data used for training the recognition model is not sufficient, we learn from the experience to synthesize images to fit the CRNN model better. In this way, we recognize the multi-length text of aviation wires in images and avoid tedious operations in the process of text recognition, such as binarization, character segmentation and character classification and so on.

3. Dataset of aviation wire text

3.1. Image acquisition for aviation wire dataset

The unusual size of the text printed on the surface of aviation wires is 46mm*1mm, as shown in figure 1 (a). The ordinary commercial digital camera cannot capture high-quality images of this kind of texts. The machine vision camera is used to capture images of aviation wires, which has higher resolution, transmission stability and anti-interference ability compared with commercial digital cameras. The field of view of the camera is determined as 60mm*45mm taking into account the maximum size of
the aviation wire text (e.g., different sizes such as 46mm*1mm, 44mm*2mm, and 20mm*3mm etc.). The working distance from cameras to the aviation wires is set from 80mm to 110mm. The precision of images captured by cameras is set to 0.2mm in view of these extreme sizes of the aviation wire text. The resolution of the image is given by

\[ \text{resolution} = \frac{\text{FOV}}{\text{Pixel}} \]

(1)

Where, FOV is the field of view for cameras, pixel is the precision of image quality.

In practice, five pixels are used to represent one resolution value in images to prevent the influence of noise. We improve the precision of images by doubling the number of pixels of images to confront the low-imaging effect caused by the uncertainty working environment. Therefore, the number of pixels on the long side of images is determined to be 3072, and the short side is 2048. Based on these requirements, the type LBAS-GE60-17M machine vision camera and FA0801C lens are selected to capture images of aviation wires, as shown in figure 1 (b). When taking images of aviation wires, it is necessary to show influences of the complicated environmental background, the variation of the light intensity and the motion blur on the image quality. The aviation wire dataset comprising of 2474 training images and 762 testing images is obtained finally. Each image has English text and Arabic numbers in it, as shown in figure 1 (c-f). Then we annotate labels for these images following the format of ICDAR 2015 [5]. It is noted that most aviation wire texts have more than nine characters and the number of the characters varies in a wide range, as shown in figure 2.

![Figure 2. Number statistics of aviation wire texts](image)

3.2. Synthetic wire dataset and augmented wire dataset

The amount of images collected manually in the workshop is not enough to train some deep neural network. Therefore, synthetic image techniques are employed to generate images having similar features with the acquired images in-site in order to obtain sufficient training data.

![Figure 3. Word templates of aviation wires and synthetic images](image)

A dictionary is formed according to the content of aviation wire texts. Each word in the dictionary is the serial number of one aviation wire. The first one is to randomly taking a certain number of words from the dictionary to generate word templates, as illustrated in figure 3 (a-c) shows some word templates with left-to-right characters while figure 3 (d-f) shows those with head-to-tail style. The red boxes in the word templates represent the position of each character in images. Then the word
templates are rendered into actual images of aviation wires. It is convenient to obtain the coordinates of the rendered word in the synthetic images according to the rendering process [7]. Synthetic images generated by this method are more natural and similar to the real images since its characters are arranged in the same style of aviation wires, as shown in figure 3 (g). The distance between each character of synthetic text is adjusted flexibly. The method is used to generate 10569 synthetic images. All these synthetic images are used to be trained for the aviation wire detection model.

Figure 4. Augmented image process

Augmented images is particular important to improve the localization accuracy of the text detection model. Here the image augmentation technique is applied on the images whose number of characters exceeds nine and the ration of length over width is greater than five, to sample repeatedly text segments of aviation wire text [2], as shown in figure 4. In the end, 8095 augmented images are sampled in the basis of the training set of aviation wire dataset. And all these augmented images are used to be trained for the aviation wire detection model.

3.3. Serial number dataset and augmented number dataset

Serial number dataset can be acquired by segmenting text region in images. A rotated rectangle (pink box) can be produced by covering the annotated quadrangle (blue box) of text regions with minimal area, as shown in figure 5 (b). Then a horizontal rectangle can be produced by rotating the image at the appropriate angle, see figure 5 (b). This contiguous rectangle region is cropped from the image as a separate image. Some cropped images in this dataset are shown as in figure 5 (c).

Figure 5. Cropped text region process

For the serial number dataset, the information of the aviation wire text is annotated with the above-mentioned format of ICDAR 2015 for recognition tasks. After removing the illegible cropped images, 58964 cropped images are acquired from the aviation wire dataset and synthetic wire dataset, 690 cropped images of those from testing set in the aviation wire dataset are used for testing, and the remainder of the text dataset are utilized to be trained the CRNN model.

Five different ways of data augmentation are adopted on 2055 cropped images from the training set in the serial number dataset to improve the recognizing accuracy. The original RGB image in serial number dataset is shown in figure 6 (a). Figure 6 (b) is the augmented image by randomly shuffling RGB channels based on the original RGB image. Figure 6 (c-f) shows the augmented image, respectively, after the process of white balance, by adding snowflakes noise, by randomly changing...
lightness, and after contrast adjustment. In the end, 10271 augmented word images are produced and all these images are used to train the recognition model.

![Image augmentation for serial number dataset](image)

Figure 6. Image augmentation for serial number dataset

4. Methodology

We propose a novel text reading technique for aviation wire assembly, which exploits deep neural network to accurately detect and recognize the text printed on the surface of aviation wires.

4.1. Flowchart for reading aviation wire text in image

A high-level overview of our flowchart is illustrated in figure 7, including the steps of capturing aviation wire images, text detection, text recognition and output.

![Flowchart for reading aviation wire text in images](image)

Figure 7. Flowchart for reading aviation wire text in images

Images of aviation wires are captured first by the machine vision camera in assembly workshop. The image is fed into the trained detection model to output series predicted feature maps after pre-processing. The bounding box of the text in image is derived from the output of the feature maps. A rectangle text region is produced by covering the bounding box with minimal area. This contiguous rectangle region, which contains serial number of the aviation wire in images, is cropped from image and then undergoes preprocessing including resizing and pixel padding. This region is then sent to the trained recognition model to predict the character sequence. This approach finally outputs the location of aviation wire text in images, and the identified text content.

In addition, the text detection model is trained by aviation wire dataset, synthetic wire dataset and augmented wire dataset. The images of these datasets are labeled with text inner-instance, text border sections and vertex offset, before they are fed into the detection model. The text recognition model is trained on serial number dataset and augmented number dataset. We detail the above steps in the following sections.
4.2. Label generation for detection objectives

Define a quadrangle \( Q_j = \{p_{i,j}\}_{i \in \{1,2,3,4\}} \), where \( p_{i,j} = \{x_{i,j}, y_{i,j}\} \) are vertices of the quadrangle in clockwise order, \( D(p_x, p_y) \) is the \( L_2 \) distance between \( p_x \) and \( p_y \). The minimum length \( r_{\text{min},j} \) for \( Q_j \) is computed as 
\[
    r_{\text{min},j} = \min\left(D(p_{i,j}, p_{(i+2) \mod 4,j}), D(p_{i,j}, p_{(i+3) \mod 4,j})\right).
\]

A text inner-instance (yellow solid box) is extracted for each image, which is a shrunk version of the original annotation quadrangle \( Q_M \) (red dashed box), as illustrated in figure 8 (a). The text inner-instance \( Q_s \) is produced by moving each edge of the \( Q_M \) inward 0.1\( r_{\text{min},M} \), see figure 8 (b), by which way, it filters the background noise of the marginal area in text regions and better represents the features of aviation wire texts.

![Figure 8. Label generation and vertex offset.](image)

- (a) Original and shrunk quadrangle of annotation.
- (b) Text inner-instance.
- (c) Short text border sections.
- (d) Long text border sections.
- (e) Horizontal shift and vertical shift.
- (f) Four channels of vertex offset.

An extended quadrangle \( Q_{L1} \) is produced by extend the two shorter edges of \( Q_S \) outward 0.2\( r_{\text{min},S} \). Then the shrunk quadrangle \( Q_{L2} \) is generated by shrinking the two shorter edges of the \( Q_S \) with the same threshold. A pair of long text border sections is generated by removing the overlaps between the \( Q_{L1} \) and \( Q_{L2} \), as shown figure 8 (d). In the same way, an extended quadrangle \( Q_{L3} \) is generated by moving the two longer edges of the \( Q_{L2} \) outward with 0.5\( r_{\text{min},L2} \). A shrunk quadrangle \( Q_{L4} \) is generated with the same threshold. A pair of short text border sections are obtained by removing the overlaps between the \( Q_{L3} \) and \( Q_{L4} \), as illustrated figure 8 (c). So the text border sections have four channels and it is generated based on the shrunk version of the original quadrangle \( Q_S \).

A 4-channel vertex offset is coordinate shift from each pixel located in the short text border sections to the closer two vertices of the annotated original quadrangle \( Q_M \), including horizontal shift and vertical shift. The generation for vertex offset is illustrated in figure 8 (e-f). The short text border sections that are closer to the vertices of annotated quadrangles is utilized to carry the information of coordinate offset, to improve the location accuracy of bounding boxes. All those label generation and normalized images are fed to the network to train the detection model.

4.3. Network architecture and training objectives for aviation wire text detection

The network architecture of our proposed detection model uses features of different levels to detect the texts of aviation wires in various sizes, as shown in figure 9. The model consists feature extractor stem, the feature-merging branch, and the output layer. The backbone of feature extractor stem in the network is VGG16 [10], which has deeper layers than PVANet [6]. We exploit the layers of pooling-2 to pooling-5 to extract feature maps of the input images. The output layer produces three per-pixel prediction maps including a text score map, a text border map, and a vertex regressor. Text score map is the confidence score of each pixel being a text pixel. Text border map is a 4-channel confidence score of each pixel being the border sections of text. Vertex regressor contains four channels to predict the coordinate shift from closer two corner vertices \( (p_i, p_j)_{i \neq j} \) of the quadrangle \( Q = \{p_k\}_{k \in \{1,2,3,4\}} \) to the pixel location. Values of four channels denote the four offsets \( (\Delta x_i, \Delta y_i, \Delta x_j, \Delta y_j) \).
The training objectives aims to minimize the formulated loss function
\[ L = L_S + \lambda_b L_b + \lambda_g L_g \]  \tag{2}
where, \( L_S \), \( L_b \) and \( L_g \) denote the loss of text score map, text border map and vertex regressor. Hyperparameters \( \lambda_b \) and \( \lambda_g \) are empirically set at 1.0 in our experiment. For the text score map loss \( L_S \), we adopt the class-balanced cross-entropy, given by
\[ L_S = -\beta \hat{Y} \log \hat{Y} - (1 - \beta)(1 - \hat{Y}) \log(1 - \hat{Y}), \beta = 1 - \frac{\sum_{y} \hat{Y}}{\sum_{y} \hat{Y}} \]  \tag{3}
where, \( \hat{Y} \) is the prediction of the text score map, and \( \hat{Y} \) is the ground truth from the text inner-instance \( Q_S \).

For the text border map loss \( L_b \), class-balanced cross-entropy is also used as the previous step. This time, the definition \( \hat{Y} \) is the prediction of the 4-channel text border map, and \( \hat{Y} \) is the ground truth from 4-channel text border sections.

For the regression losses \( L_g \), we use the smoothed-L1 loss in training, computed as
\[ L_g = \text{smoothed}_{\ell_1} (\hat{c}_i - \tilde{c}_i) \]  \tag{4}
where, \( \hat{c}_i \) is the prediction of the 4-channel vertex regressor, and \( \tilde{c}_i \) is the ground truth from 4-channel vertex offset.

**Figure 9. Network Architecture of aviation wire text detection**

**4.4. Post-processing to generate candidate aviation wire text**

The detection model outputs a text score map, a text border map, and a vertex regressor. Based on the two former map, a number of rough-text blocks (orange colors), short text border blocks (yellow and red colors) and long text border blocks (green and blue colors) are determined by global thresholding where the threshold is empirically set at 0.8, see figure 10 (a). As shown in figure 10 (b), the inner-text blocks (orange colors) are delineated by removing the overlaps between the rough-text blocks and two long text border blocks. Overlaps between each inner-text block and two short text border blocks are regarded as side-pixels, including head-pixels (yellow colors) and tail-pixels (red colors).

For the 4-channel vertex regressor \( (\Delta x_i, \Delta y_i, \Delta x_j, \Delta y_j) \), pixels overlap with side-pixels are determined position-pixels, including left-side position-pixels (overlaps with head-pixels) and right-side position-pixels (overlaps with tail-pixels). Each pixel of left-side position-pixels is used to predict the two left corner vertices of the bounding box, and the remainder pixels are used to predict the other two corner vertices. The coordinates of corner vertices of the bounding box are obtained by
\[ \hat{p}_i = \left( \frac{\sum_j x_j w_j / \sum_j w_j, \sum_j y_j w_j / \sum_j y_j w_j}{} \right) \]  \tag{5}
where, $\bar{p}_i$ is the coordinate of four corner vertices of the bounding box, $\chi_j$ is pixel value of the position-pixels, $\omega_j$ is the corresponding confidence value of the side-pixels, $n$ is the number of pixels in the position-pixels region. In this way, it reduces computing time to obtain bounding boxes compared with that uses non maximum suppression (NMS) [8].

4.5. Aviation wire text region refinement and recognition

Text regions in the image are cropped from the images according to the bounding boxes obtained after post-processing steps. The cropped text region is resized first by setting the height of this image to 64 pixels and the width to $N \times 16 -16$, where $N$ is an integer derived reversely from recognition network CRNN to prevent the calculation process of convolution and pooling in CRNN model from crossing the boundary of images, given by

$$N = \max[2, (w \times 4/h + 1.5)]$$

(6)

where, $w$ is the width and $h$ is the height of images before resizing. To ensure the text locate in the centre of the cropped image, 32 blank pixels are padded to two sides of the image. The final image of text region has a height of 64 pixels and width of $N \times 16 + 48$ pixels. The above region refinement is shown in figure 11. The CRNN model is employed as the text recognizer and the output text content is obtained by putting the cropped image after region preprocessing to the model.

5. Experiments

In this section, we verify the effectiveness of the proposed algorithm on two different tasks, including text detection and text recognition. To compare the proposed approach with existing methods, we conducted experiments on aviation wire dataset and serial number dataset respectively. More details about these datasets have been discussed in the Part 3.

5.1. Evaluation criterion and implement details

For now, there are not a unified criterion to evaluate the performance of aviation wire detection and recognition. In this paper, we adopt the evaluation criterion for scene text detection task in ICDAR 2015. We mainly evaluate the detection model for aviation wires in case that the overlap (IoU) between the detection and ground truth bounding box is greater than 0.5, i.e., 0.5, 0.6 and 0.7.

| IoU | 0.4 | 0.5 | 0.6 | 0.7 |
|-----|-----|-----|-----|-----|
| Method | P | R | F | P | R | F | P | R | F | P | R | F |
| East | 33.33 | 65.28 | 44.13 | 8.63 | 16.91 | 11.43 | 6.11 | 11.96 | 8.09 | - | - | - |
| AdvancedEAST | - | - | - | 96.26 | 97.01 | 96.64 | 93.81 | 94.55 | 94.18 | 87.76 | 88.44 | 88.10 |
| Ours | - | - | - | 97.55 | 98.44 | 97.99 | 96.78 | 97.66 | 97.22 | 92.14 | 92.98 | 92.56 |

As for aviation wire text recognition, a correctly recognized aviation wire text means all the characters of the text are recognized correctly (i.e., word recognition). In addition, we also provide the accuracy of character recognition, which is defined as the number of correctly recognized characters
divided by the total number of characters from the ground truth. Following the criterion of ICDAR 2015, the total normalized edit distance (NED) is also provided on all the test images.

The detection network is trained with 736*736 images optimizing by Adam optimizer with starting learning rate of 10^{-3} and batch size of 4. Based on the VGG16 model pre-trained on the ImageNet data [11], we initialize the new layers (e.g., the feature-merging branch and output layer) by using random weights with Gaussian distribution of 0 mean and 0.01 standard deviation. The recognition network is fine-tuned from a pre-trained CRNN model.

5.2. Aviation wire text detection
The detection performance of our detection model is shown in Table 1 for the testing images of aviation wire dataset. Based on the evaluation criterion described above, our approach outperforms all the three methods in both precision and recall. To be specific, it achieves 98.44, 97.55 and 97.99 in recall, precision and F-score, which are under the IoU of 0.5. Some previous approaches are also evaluated for comparison. The detection accuracy significantly of EAST declines when the IoU value is set more than 0.5. The AdvancedEAST [12] outputs competitive performance with the IoU is set to 0.5. However, the performance gap between our model and AdvancedEAST appear conspicuous as the value of IoU increases and our model shows a prominent advantage.

Figure 12 shows several sample images of the corresponding detections by using the proposed model. From experience, the two short text borders play an important role in detecting long text with high localization accuracy, such as figure 12 (b). The two long text borders have capacity to successfully distinguish the closely adjacent text blocks, as shown in figure 12 (c).

5.3. Aviation wire text recognition
The trained recognition model reaches an accuracy of character recognition of 91.0595%. In term of the accuracy of word recognition, it achieves an accuracy of 73.9130%. In addition, the trained model achieves a value of 309 for NED in the testing images. Some positive examples of recognition results of aviation wires are shown in figure 12. A few negative examples of incorrect recognition results are given as figure 12 (g-i).

One of the important components in recognition model is CTC, which is also prone to convert the repeated characters to one character. Therefore, for the aviation wire text composed of a series repeated characters, it is more difficult to correctly recognize the text.
6. Conclusion
In this work, we propose a deep network model for detecting and recognizing the aviation wire texts in the images gathered under the real environmental background of assembly workshops. The detection model uses the features of the text border sections as the training objectives, it effectively detects the texts with large character spacing. The strategy that utilizing pixels in the text border sections to regress the two closer coordinates of bounding boxes, improves the location accuracy of text regions. Then extra pixels are padded to the text region cropped from the image. The CRNN is capable of recognizing the text region with various numbers of characters. Future work will focus on improving the accuracy of the recognition model. In addition, integrating the proposed approach of aviation wire text reading into practice to improve the efficiency of aviation wire assembly is also interested.

Acknowledgment
This work is funded by the Graduate Innovation Fund of Beihang University, and the program of Smart Assistant Assembly based on VR/AR Technology (MJ-2017-G-70).

References
[1] Zhou X, Yao C, Wen H, Wang Y, Zhou S, He W and Liang J 2017 EAST: an efficient and accurate scene text detector Proc.of the IEEE Conf. on Computer Vision and Pattern Recognition 2642-51
[2] Xue C, Lu S and Zhan F 2018 Accurate Scene Text Detection Through Border Semantics Awareness and Bootstrapping Proc. of the European Conf. on Computer Vision 370-87
[3] Jaderberg M, Simonyan K, Vedaldi A and Zisserman A 2016 Reading text in the wild with convolutional neural networks Int J ComputVision 116(1) 1-20
[4] Shi B, Bai X and Yao C 2017 An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition IEEE T Pattern Ana 39(11) 2298-304
[5] Karatzas D et al. 2015 ICDAR 2015 competition on Robust Reading 2015 13th Int. Conf. on Document Analysis and Recognition, IEEE Computer Society 1156-60
[6] Kim K, Hong S, Roh B, Cheon Y and Park M 2016 PVANET: Deep but Lightweight Neural Networks for Real-time Object Detection Preprint arXiv/1608.08021
[7] Gupta A, Vedaldi A and Zisserman A 2016 Synthetic data for text localisation in natural images Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition 2315-24
[8] Neubeck A and Gool L 2006 Efficient Non-Maximum Suppressio 18th Int. Conf. on Pattern Recognition 850-5
[9] Graves A, Fernández S, Gomez F and Schmidhuber J 2006 Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks Proc. of the 23rd Int. Conf. on Machine learning. ACM 369-76
[10] Simonyan K and Zisserman A 2014 Very deep convolutional networks for large-scale image recognition Preprint arXiv/1409.1556
[11] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S and Huang Z 2015 ImageNet Large Scale Visual Recognition Challenge Int J ComputVisio 115(3) 211-52
[12] https://github.com/huoyijie/AdvancedEAST