Research Article

Semantic Segmentation of PHT Based on Improved DeeplabV3+

Haiquan Fang

Zhejiang University of Technology, Hangzhou 310023, China

Correspondence should be addressed to Haiquan Fang; fanghaiquan@zjut.edu.cn

Received 9 February 2022; Revised 27 February 2022; Accepted 3 March 2022; Published 19 March 2022

Academic Editor: Yuxing Li

Copyright © 2022 Haiquan Fang. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This work aimed to address the two shortcomings of the printed and handwritten texts (PHT) classification. The classification accuracy of FCN and U-net, which are used for PHT pixel-level classification, still has room to improve. PHT public datasets have small sample sizes, and the generalization ability of the models is not good. In this paper, first, a pixel-level sample-making method for PHT identification was proposed, and a PHT dataset 2021 (PHTD 2021), containing 3,000 samples, was constructed. Second, because there is a large number of words but the contours are small in documents, the DeeplabV3+ model was improved. The network layer number and pooling times were reduced, and the convolution kernel and dilated rate were increased. In the experiment, the improved DeeplabV3+ model had a classification accuracy of 95.06% on the test samples from the PHTD 2021 dataset. The improved DeeplabV3+ model has a higher recognition accuracy than the FCN and DeeplabV3+ models. Finally, after the classification of PHT, applications of handwritten texts removal and handwritten texts extraction are provided.

1. Introduction

Information and communication technologies have developed rapidly and spread worldwide. In particular, the deepening application of big data, artificial intelligence, cloud computing, Internet of things, and mobile Internet, which provides unlimited possibilities for an intelligent society. Digitalization and intellectualization are the general trends of social development. Paper documents are closely related to people's lives, studies, and work. The digitization of document images is convenient for people. It is an important aspect of the development of social intelligence. When it comes to the digitization of document images, people usually think of optical character recognition (OCR). When the document image contains both printed and handwritten text (PHT), people may have new digital requirements, such as the character recognition of PHT, respectively, removing handwritten characters, retaining only printed characters, and extracting handwritten characters. In order to meet these requirements, the classification of PHT is a key problem that must be resolved.

How is a document picture containing both PHT accurately classified? This question has not been well answered at present. If we can accurately classify PHT, then we can separate them. Using character recognition technology, we can recognize PHT, respectively. Therefore, this technology has a wide range of application prospects in digital reform, including in education, finance, and government affairs. Typical application scenarios include the automatic marking of test papers and bank bills, as well as the digitization of personal and professional files. The existing character recognition technology recognizes PHT together, but it cannot distinguish between them. Therefore, the classification of PHT has important research value.

Here, the classification of PHT was used as a semantic segmentation problem in the field of artificial intelligence image processing. An advanced semantic segmentation algorithm was adopted to achieve a higher classification accuracy. There are three innovations in this research. First, a pixel-level sample-making method for PHT was proposed; second, the DeeplabV3+ model was improved, allowing it to classify PHT with high accuracy; and third, further applications after the classification of PHT, including handwritten text removal and handwritten text extraction, are provided.

This research contributes to the research field by accurately classifying PHT, and it can play a role in promoting the digitization of document images containing both PHT.
After classification, more digital functions can be realized. For example, (1) PHT can be separated, and then, in combination with OCR, PHT can be recognized, respectively. (2) The handwritten texts are removed and the maintenance of only the printed texts are carried out. (3) The handwritten texts are extracted. These functions have important applications in the digital reform of education, finance, and government affairs.

2. Related Work

The early classification methods of PHT mainly include the hidden Markov model [1], discriminant analysis [2], multilayer perceptron model [3], wavelet transform, and support vector machine model [4]. The classification accuracy of these methods for PHT is not high, and it is difficult to distinguish the overlapping parts of PHT.

In recent years, with the rapid development of artificial intelligence, deep learning methods have been widely used in many fields. Deep learning methods can also be used for the classification of the PHT. Literature [5] applied fully convolutional networks (FCN) [6] to classify the PHT. The classification accuracy of FCN is higher than that of traditional methods, and it can also distinguish the overlapping parts of PHT. However, owing to a sample size of only 80, the generalization ability of the model is weak. Literature [7] applied U-Net [8] to separate handwritten texts, and the dataset used in [7] was obtained by synthesis; thus, it did not represent a real sample. The two methods are early classical algorithms for semantic segmentation.

At present, the classification of PHT has not been well addressed for two main reasons. First, the sample size of the PHT in the public dataset is too small, and second, the algorithm model needs to be improved.

Public datasets in the field of computer vision, such as the VOC and COCO datasets, were constructed by manual labeling. The objects in these datasets are large, and manual labeling is relatively easy. However, for document pictures containing both PHT, owing to the large number of words and the small contours, manual labeling is time-consuming and inaccurate. The production of PHT datasets is a difficult problem. Therefore, our research proposed a new dataset-making method that can compile PHT datasets quickly and accurately.

The classification of PHT requires the pixel-level classification of printed texts, handwritten texts, and background, which may be addressed using semantic segmentation from the field of computer vision. In recent years, many advanced semantic segmentation algorithms have been proposed. For example, FCN [6] and U-Net [8]. The main idea behind FCN and U-Net networks is the adoption of convolution and deconvolution structures to realize pixel-level classifications. The two networks take input of arbitrary size and produce a correspondingly-sized output. The DeeplabV3+ [9] algorithm has the advantages of the Deeplab series [10–12] methods and has many improvements compared with the FCN and U-Net algorithms, making it a very advanced algorithm in the field of semantic segmentation.

The DeeplabV3+ algorithm integrates the advantages of depthwise separable convolution [13–15], residual network [16], dilated convolution [17–20], spatial pyramid pooling (SPP) [21], and atrous spatial pyramid pooling (ASPP) [11] modules. Depthwise separable convolution, factorizing a standard convolution into a depthwise convolution, followed by a pointwise convolution, drastically reduces computation complexity. Residual networks can ease the training of networks that are substantially deeper. Dilated convolutions support the exponential expansion of the receptive field and improve network performance. SPP and ASPP probe convolutional features at multiple scales. However, the DeeplabV3+ structure is complex and includes time-consuming calculations. For the classification of PHT, because the number of characters is large and the contour is small, we maintained the advantages of the DeeplabV3+ model and improved it. The network layer numbers and pooling times were reduced, and the convolution kernel and dilated rate were increased. Reducing the number of network layers increased the computing speed, and the pooling times were reduced to increase the resolution. The convolution kernel and dilated rate were increased to obtain a larger receptive field, which improved the classification accuracy.

The classification of PHT represents one aspect of research on document images. Research on document images also includes text segmentation [22], OCR [23, 24], stamp extraction [25], and document classification [26]. In [22], a novel text segmentation approach, TexTR-Net, was proposed, and it improved text segmentation performance by nearly 2% compared with other methods. In [24], a practical ultra-lightweight OCR system, PP-OCRv2, was proposed, and the precision of PP-OCRv2 was 7% higher than that of PP-OCR [23]. Text segmentation and OCR do not distinguish between PHT. Reference [25] proposed a lightweight model for passport stamp detection and classification. The model used contour detection and texture feature extraction, and then the stamps were divided into three categories by using a multilayer perceptron, entry local stamp, exit local stamp, and foreign stamp. The accuracy achieved was 0.945. Reference [26] applied a deep convolution neural network to classify documents and divided them into six categories: advertisement, e-mail, form, memo, news, letter, note, report, resume, and scientific, with an accuracy of 0.931.

From the view of earlier research, many artificial intelligence technologies have been applied to the digital research of document images. However, there are still some limitations in the classification of PHT, which are mainly reflected by the low classification accuracy. These problems occur because there are not enough training samples and the applied algorithm is not advanced enough.

3. Methods

DeeplabV3+ performs well in semantic segmentation, but the network structure is complex and a long computing time is needed. In addition, DeeplabV3+ was designed for VOC datasets. The objects in the VOC dataset are large, and the number is small. But the objects in the PHT dataset are small,
and the number is large. Therefore, to better classify PHT, DeeplabV3+ was improved. The improved DeeplabV3+ still retains the encoder-decoder structure. The encoder module includes Xception and ASPP. The improvements were mainly as follows:

1. Reduced the number of layers in the network. The Xception of DeeplabV3+ includes three parts: entry flow, middle flow, and exit flow. Here, we maintained the entry flow and omitted the middle flow and exit flow.

2. Deeplabv3+ has for a use of 5 times stride, which is the equivalent to 5 pooling. We used pooling only 2 times. Because the object contours of the PHT dataset are small, pooling too many times will reduce the resolution.

3. The convolution kernel used for Xception in DeeplabV3+ is $3 \times 3$, and the dilated rate is 1. Our convolution kernel is $9 \times 9$, the dilated rate is 2 or 3, which resulted in a larger receptive field.

4. Reduced the number of channels in the convolution kernel, which reduced the amount of calculations required.

The improved DeeplabV3+ network structure is shown in Figure 1. The details of the model are as follows:

1. Input layer: the input figure is $512 \times 512$, with 3 channels.

2. Dilated convolution layer: there are 32 channels, the kernel is $9 \times 9$, dilation is 2, and stride is 1.

3. ResNet module 1
   - (3.1) Depthwise separable convolution: there are 64 channels, the kernel is $9 \times 9$, dilation is 3, and stride is 1.
   - (3.2) Depthwise separable convolution: there are 64 channels, the kernel is $9 \times 9$, dilation is 3, and stride is 1.
   - (3.3) Depthwise separable convolution: there are 64 channels, the kernel is $9 \times 9$, dilation is 3, and stride is 1.
   - (3.4) Pool layer: the down-sample ratio is $2 \times 2$, and the result of pooling is recorded as residual3.
   - (3.5) Convoluted the results obtained in (2), there are 64 channels, the kernel is $1 \times 1$, dilation is 1, and stride is 2. The result is recorded as shortcut3.
   - (3.6) residual3+shortcut3

4. ResNet module 2
   - (4.1) Depthwise separable convolution: there are 128 channels, the kernel is $9 \times 9$, dilation is 3, and stride is 1.
   - (4.2) Depthwise separable convolution: there are 128 channels, the kernel is $9 \times 9$, dilation is 3, and stride is 1.

4.4 Pool layer: down-sample ratio is $2 \times 2$, and the result of pooling is recorded as residual4.

4.5 Convoluted the results obtained in (3.6), there are 128 channels, the kernel is $1 \times 1$, dilation is 1, and stride is 2. The result is recorded as shortcut4.

4.6 residual4+shortcut4

5. ASPP module
   - (5.1) ASPP branch 1 is convolution: there are 64 channels, the kernel is $1 \times 1$, dilation is 1, and stride is 1.
   - (5.2) ASPP branch 2 is depthwise separable convolution: there are 64 channels, the kernel is $3 \times 3$, dilation is 4, and stride is 1.
   - (5.3) ASPP branch 3 is depthwise separable convolution: there are 64 channels, the kernel is $3 \times 3$, dilation is 8, and stride is 1.
   - (5.4) ASPP branch 4 is depthwise separable convolution: there are 64 channels, the kernel is $3 \times 3$, dilation is 16, and stride is 1.
   - (5.5) ASPP branch 5 is global average pooling. Then, resize to the same size as the result of (4.6), with a $1 \times 1$ convolution layer and 64 channels.

5.6 Concatenate 5 ASPP branches

6. Convolution layer: there are 64 channels, and the kernel is $1 \times 1$.

7. Resize the result of (6) to the same size as the skip, and the up-sample ratio is $2 \times 2$.

8. The skip is convoluted with 32 channels and the kernel is $1 \times 1$.

9. Concatenate the results of (7) and (8)

10. Depthwise separable convolution layer: there are 128 channels, the kernel is $3 \times 3$, dilation is 1, and stride is 1.

11. Depthwise separable convolution layer: there are 128 channels, the kernel is $3 \times 3$, dilation is 1, and stride is 1.

12. Convolution layer: there are 3 channels, and the kernel is $1 \times 1$.

13. Resize the result of (12) to the same size as the input picture, and the up-sample ratio is $2 \times 2$.

4. Experiment and Application

4.1. Technical Route. The technical route of this research is shown in Figure 2. It includes model training, detection, and application. Model training was as follows: making PHT datasets, establishing an improved DeeplabV3+ model, inputting the training samples into the model for training, and
Figure 1: Network structure of the proposed method.

Figure 2: Technological roadmap.
producing the trained model. Model detection was as follows: inputting the test samples into the trained model for detection, calculating the classification accuracy of the model, and evaluating the performance of the model. If the performance is good, then the model can be applied. Model application was as follows: taking photos or scanning the documents containing both PHT to obtain images, preprocessing the image, including clipping and adaptive binarization into black-and-white images, and then inputting them into the trained model detection system to obtain the resulting classified image.

4.2. Dataset Construction Method. Owing to a lack of datasets containing both PHT in pictures, a method to construct a pixel-level dataset of PHT was proposed. The conventional data labeling methods (such as labelme) require manual labeling on the computer. The dataset-making method proposed in this paper is automatically processed by an algorithm program without manual annotation. Therefore, the produced samples have high annotation accuracy, and the method saves both time and labor.

The steps are as follows:

Step 1. we select a piece of A4-size black-and-white paper having a printed font. The printed font is black and the background is white.

Step 2. we write on the paper with a red ink pen.

Step 3. after writing, we scan with a scanner to produce a picture, as shown in Figure 3(a).

Step 4. the image shown in Figure 3(a) is preprocessed using the algorithm program, including binarization into a black-and-white figure. The results obtained are shown in Figure 3(b). This is the input sample of the training model. The PHT in Figure 3(b) is black, and the background is white.

Step 5. we classify the printed texts, handwritten texts, and background in Figure 3(a) at the pixel level using the algorithm program. The three categories are one-hot coded. The printed texts are represented by [1, 0, 0], the handwritten texts are represented by [0, 1, 0], and the background are represented by [0, 0, 1]. The result is recorded as a mask matrix, which contains the output sample of the training model and the label data of the sample. To visually see the classification effect, different colors are used instead of different categories. The printed texts are represented by blue [255, 0, 0], the handwritten texts are represented by green [0, 255, 0], and the background is represented by white [255, 255, 255]. The results are shown in Figure 3(c). The classification principle is based on different colors corresponding to different pixel values. Because of the different colors, the pixel values corresponding to the red handwritten text, black printed text, and white background are quite different and can be easily distinguished using the algorithm program.

We used the above steps to construct the dataset, which contained 3,000 samples, including 2,000 training, 300 evaluation, and 700 test samples. A sample here refers to a photo of a piece of A4 paper scanned by the scanner. The materials used to make samples were selected from papers on various subjects in primary and secondary schools. The languages of the PHT included English and Chinese. In total, 10 people provided handwriting.

This dataset was named PHT dataset 2021 (PHTD 2021). All the models were trained, validated, and tested on the dataset PHTD 2021. The distribution of dataset PHTD2021 is shown in Table 1.

The resolution of each picture in dataset PHTD2021 is approximately 3,000 × 2,000. Before training the model, each picture was randomly cropped into 20 small pictures, and the size of the small picture was 512 × 512. Thus, the training samples of the input model were 2,000 × 20 = 40,000.

For the image T, suppose the length of the image is a1 and the width is b1. Thus, the number of rows in matrix T is b1 and the number of columns is a1. Then, a point (x, y) was randomly selected. x and y are random numbers generated by a computer program. The selection range of x is [0, a1-w], and the selection range of y is [0, b1-h]. Then t1 = T[y: y+h, x: x+w] is a small image with a length of w and a height of h randomly cropped from image T. Here, h = w = 512. The input sample and output sample were cropped synchronously. Thus, the two images used the same location point (x, y). In this way, we ensured that the input sample and output sample corresponded to each other.

4.3. Model Training. For experimental comparison, FCN, DeeplabV3+, and improved DeeplabV3+ models were selected.

Based on the deep learning frameworks of TensorFlow and Keras, an improved DeeplabV3+ model was constructed. The hardware configuration was as follows: NVIDIA RTX 3060 Ti graphics card, 12 GB memory, and i9-10900k CPU. The computer’s operating system was Linux.

4.4. Evaluation Index. The experimental results use the mean intersection over union (mIoU) as the evaluation standard.

\[
\text{IoU} = \frac{TP}{TP + FN + FP}
\]  

(1)

Here, TP represents the number of pixels predicted correctly in this category, FN represents the number of pixels belonging to this category but not predicted correctly, and FP represents the number of pixels not belonging to this category but predicted as this category. mIoU is obtained by calculating the average value of the IoU of all test samples. The result of calculating the average value of mIoU of printed texts, handwritten texts, and background is recorded as the total mIoU.

4.5. Experimental Results. After the three models were trained, they were each tested on the test set, and the test results are shown in Table 2. The improved DeeplabV3+
performed better than the FCN and DeeplabV3+ models. The total mIoU of the FCN, DeeplabV3+, and improved DeeplabV3+ models was 78.63%, 87.78%, and 95.06%, respectively.

After the improved DeeplabV3+ model was trained, a test sample Figure 4(a) was classified. The classification result is shown in Figure 4(b). The handwritten texts, printed texts, and background are represented by green, blue, and white, respectively. The classification effect was good. The overlapping parts of PHT were classified accurately. There are two reasons why the improved DeeplabV3+ method achieved high classification accuracy for PHT. First, this is due to the large sample size in the dataset; and second, the improved DeeplabV3+ model performs better than previous models.

4.6. Optimization of Model Hyperparameter Values. For the improved DeeplabV3+ model, the selected hyperparameter values were determined using many experiments. For example, for the selection of the convolution kernel size of Xception, consider 3 × 3, 5 × 5, 7 × 7, 9 × 9, and 11 × 11 in these cases, while other parameters of the model remain unchanged. After training, test samples were used to obtain the classification accuracy, as shown in Table 3. The larger the convolution kernel, the higher the classification accuracy. However, the larger the convolution kernel, the longer the calculation time. Therefore, the convolution kernel size should not be set too large. Considered comprehensively, a 9 × 9 convolution kernel size was appropriate.

4.7. Application. Applications for the classification of pictures containing both PHT. One application is the separation of PHT. The separation method was as follows: using the original figure of Figure 4(a) as an example. By classifying the original figure, Figure 4(b) was obtained, and then, the pixels of the handwritten text category were replaced with 255, which corresponds to white in the computer program. Consequently, the handwritten text was converted to a white background, whereas the printed text remained unchanged. The results are shown in Figure 4(c). Thus, the image in Figure 4(c) shows the printed text extracted from the original picture, which can also be regarded as the printed text obtained after removing the handwritten text. Similarly, Figure 4(b) was obtained by classifying Figure 4(a), and then, the pixels of the printed text category were replaced with 255. Consequently, the printed text was converted to a white background, whereas the handwritten text remained unchanged. The result is shown in Figure 4(d). Thus, the image in Figure 4(d) shows the handwritten text extracted from the original picture.

After the separation of pictures containing both PHT, there are more applications. For example, when combined with optical character recognition (OCR) technology, PHT can be recognized independently. The existing technology usually puts the PHT together for character recognition. In addition, after OCR, combined
with speech synthesis technology or speech cloning technology, the PHT can be converted independently into speech.

5. Conclusion

This paper realized the pixel-level classification of PHT contained in document pictures by employing the following:

A pixel-level sample-making method for PHT was proposed. Using this method, the dataset PHTD2021, containing 3,000 samples, was constructed. This dataset-making method is simple and accurate. It does not require manual annotation but is realized automatically using an algorithm program. At present, the texts in the sample include Chinese and English languages. This method can be used to make samples of any language, providing a new way to produce printed and handwritten datasets of various languages.

The DeeplabV3+ algorithm model was improved and applied to the classification of PHT. The improvement is mainly reflected in the reduction in the network layer numbers and pooling times, as well as increases in the convolution kernel and dilated rate. Owing to the large number of training samples, the trained model had good generalization and high classification abilities. The total mIoU was 95.06%, which was better than that of the FCN and DeeplabV3+ models. Additionally, this method can accurately classify the overlapping parts of PHT.

Further applications after the separation of PHT include handwritten text removal and handwritten text extraction. The results are ideal. More applications can be developed, such as combining it with OCR or speech synthesis technology.

Although the printed text is required to be black when constructing training samples, it can also be applied to color fonts because the image needs to be preprocessed before entering the model. The preprocessing includes border clipping and adaptive binarization. Preprocessing can convert a color image to a gray image, which widens the application scope and allows it to be applied to document images having various color fonts. Therefore, the classification method of printed and handwritten texts in this paper has theoretical and application values.

The method proposed in this paper has two limitations: (a) because the training samples are all scanned images, the classification effects of images taken by mobile phones will be lower than those of scanned images; and (b) because the training samples only included English and Chinese characters, the PHT of other languages cannot be well classified.

Data Availability

The raw data required to reproduce these findings cannot be shared at this time as the data also forms part of an ongoing study.

Conflicts of Interest

The author declares that there are no conflicts of interest.

References


