Automatic Identification of Bond Information Based on OCR and NLP

Jizhe Dai*, Zhengyan Ma
CFETS Information Technology (Shanghai) Co. ShangHai , China.

* Corresponding author. Email: daijizhe_zh@chinamoney.com.cn
Manuscript submitted February 10, 2019; accepted May 20, 2019.
doi: 10.17706/jcp.14.6.397-403

Abstract: With the rapid development of the financial industry, the realization of automatic recognition of bond information is a hot issue in the current financial science and technology research. This paper realizes the automatic recognition of bond information through the combination of OCR and NLP technology. The text information of bond is extracted from the picture by OCR technology. In view of various kinds of interference in the image, the difference subtraction operation based on HSV model of color image is adopted to get better image information for image-based sequence model. The image-based sequence model is composed of convolution neural network and recursive neural network. The bond information can be well recognized by the model. Finally, the text information is processed by the algorithm based on Levenshtein distance, and the key-value is extracted, which has a certain practical value.

Key words: Image processing, OCR, graph sequence model, natural language processing.

1. Introduction

With the acceleration of economic globalization and the development of the financial industry, the number of bond manuscripts produced daily is increasing month by month. At present, most of the information of the bond manuscripts is stored in picture and text format. When the information of the bond manuscripts is input, it needs to check the information manually, which wastes a lot of human resources. In order to improve the efficiency and accuracy of manuscript input, OCR character recognition technology combined with natural language processing technology is proposed for manuscript recognition.

OCR technology, also known as optical character recognition technology, is widely used in bank card recognition and certificate recognition. When OCR technology is applied to manuscript recognition, the picture information can be effectively converted into text information. OCR mainly involves image processing and character recognition. In traditional OCR, image processing mainly filters the noise of the image, recognizes the processed image, identifies the text content of the bond, and then establishes the key-value model of the bond keyword and the content through natural language processing

In conventional optical character recognition, it is necessary to train a strong character detector to detect accurately and crop each character from the original image. However, these models based on single-character recognition do not utilize the context of the text, but simply perform image recognition according to the glyphs; and these character recognitions require high-quality preprocessing, such as character segmentation. Some words are more easily and accurately identified in the sequence than in the case of a single character. For example, "孑" and "子" in "茕茕孑" are similar, but are easily distinguished in

the sequence .

2. Algorithm Model

The technology introduced in this paper mainly involves three aspects, namely, image processing technology, intelligent recognition technology and natural language processing technology.

2.1. Image Processing Technology in OCR

Image processing techniques in traditional OCR, only filtering and sharpening operations are performed on the image to be recognized. When there is a large interference in the image to be recognized, there will be a large error. As shown in Fig. 1.

Fig. 1. Image with larger interference.

In Fig. 1, there is a larger red seal interference. In the absence of interference, only a simple denoising and sharpening of the image, usually using spatial linear low-pass filtering and USM sharpening can achieve better results, but in the case of chop interference, there will be garbled when identifying, and the accuracy rate will drop sharply. In view of the above situation, the interference color in the image is faded to remove the redundant red stamp and other background colors. The color models in images are usually represented by RGB, HSV and so on. The model is shown in Fig. 2.

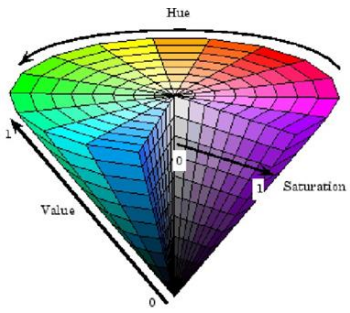

Fig. 2. (a)HSV color image model.

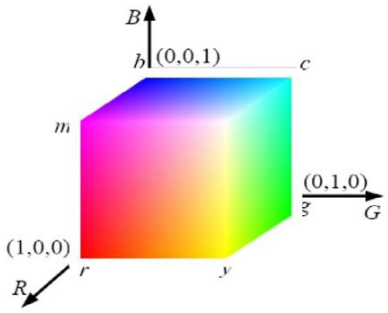

Fig. 2. (b) RGB color image model.

In the HSV model, H represents hue, which is related to the wavelength of light wave, and it represents the sensory perception of different colors; S represents saturation, which indicates the purity of the color. Pure spectral colors are completely saturated, adding white light will dilute the saturation. V represents the brightness value, which corresponds to the brightness of the image and the grayscale of the image and the brightness of the color. The coordinate system of HSV model can be cylindrical coordinate system, but it is usually represented by six pyramids.

The three axes of the RGB model represent R, G, and B respectively. As shown in Fig. 2 (b), the RGB model is a cube with black origins and white vertices farthest from the origin. RGB is based on the superposition of light. Red light plus green light plus blue light equals white light.

In this paper, HSV model is used to describe the color image, the red area and other background clutter are extracted, and the interference information of the stamp is removed by subtraction operation in the

original image. In view of the inconsistent font size in the image, bilinear interpolation is used to enlarge or reduce the smaller text, so that the font size is consistent and easy to recognize.

Bilinear interpolation is to use the pixel values of the corresponding four points in the image to determine the pixel values of the target image, using four nearest pixels to estimate the given gray level. The core idea is to perform linear interpolation in *X*-axis and *Y*-axis, assuming that the pixel value of the target is $P(x,y)$, and the four points around it are expressed as $Q_{11}(x_1,y_1)$, $Q_{12}(x_1,y_2)$, $Q_{21}(x_2,y_1)$, $Q_{22}(x_2,y_2)$.

The linear interpolation on the *X* axis is as follows,

$$f(R_2) \approx \frac{x_2-x}{x_2-x_1} f(Q_{12}) + \frac{x-x_1}{x_2-x_1} f(Q_{22}) \tag{1}$$

$$f(R_2) \approx \frac{x_2-x}{x_2-x_1} f(Q_{12}) + \frac{x-x_1}{x_2-x_1} f(Q_{22}) \tag{2}$$

The linear interpolation on the *Y* axis is as follows,

$$f(p) \approx \frac{y_2-y}{y_2-y_2} f(R_1) + \frac{y-y_1}{y_2-y_1} f(R_2) \tag{3}$$

The bilinear interpolation method is used to reduce the size of the small text, which is the same as the heading font. Then it is filtered and sharpened to remove the noise.

2.2. Text Recognition Based on Graphic Sequence Model

The problem have drawn much attention, and there have been some researches attempt to solve it. For example, one method is to first detect a single character and then use the Convolutional Neural Network (CNN) model to identify these detected characters [1], [2], while Jaderberg *et al.* assign a label to each English word to identify the text, regarding the problem as image classification [3].

Unlike general object recognition predicting a single label, identifying a sequence of such objects typically requires the recognition system to predict a series of labels. One property of sequences is that their lengths are not fixed, which can vary greatly. For example, a string may consist of 4 characters, such as "date"; it may also consist of multiple characters, such as "bond yield." Therefore, popular models with fixed out such as CNN [4], [5] cannot be directly applied to sequence prediction problems, because the model cannot produce label sequence with a variable length. Therefore, a completely CNN-based model is not suitable for direct sequence recognition.

The Recurrent Neural Network (RNN) model is another important branch of the deep neural network model, mainly used to process sequences. One advantage of RNN is that it does not require the location of each element in the sequence, no matter training or testing. The RNN model typically requires the input image to be preprocessed into an image feature sequence. For example, image feature sequence can be a set of geometric features extracted from handwritten text [6], or continuous HOG features like the method proposed by Su and Lu [7].

Considering the characteristics of CNN and RNN, we construct a model suitable for text recognition by combining the two type of models, as shown in Fig. 3. At the bottom of the model, the convolutional layers automatically extracts a sequence of features from each input image. After the convolutional layers, a recursive neural network layer is followed for predicting the frames of the feature sequence output by the convolutional layer. For the input image of the model, all images need to be preprocessed to the same height. A series of feature vectors are produced by the convolutional layer as input to the recursive layer.

Each feature unit of the sequence is generated from left to right in columns on the feature map.

The traditional RNN unit has the problem of gradient disappearance [8], which limits its ability to use context and increases the difficulty of training. Long short-term memory cells [9]-[13] (LSTM) are widely used as an improved RNN unit. Therefore the RNN unit used by the recursive layer of the model is LSTM.

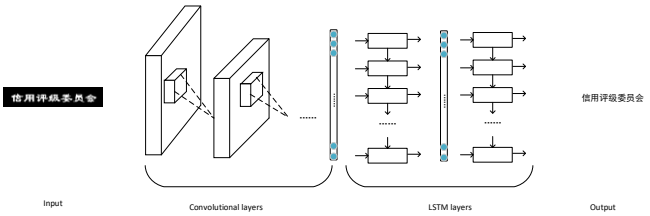

Fig. 3. OCR model structure.

2.3. Keyword Fault-Tolerance Extraction

The extraction of keywords is an important part of the automatic recognition of bond information, but the extraction of keywords becomes challenging due to the complex relationship of these keywords. The keywords may have overlap. For example, the keyword B may be a subsequence of the keyword A. For keyword A, there is a matching value Va to be matched, but the keyword B may appear between A and Va. In addition, since the OCR recognition can not be guaranteed to be completely correct, in order to extract the key information as much as possible, a keyword fault-tolerant extraction method is proposed.

The algorithm utilizes the Levenshtein distance. The Levenshtein distance is a measure used to compare the differences between two strings. In general terms, the Levenshtein distance between two words can be thought of as the minimum number of single-character edits (inserts, deletes, or replacements) required to change one string to another.

Algorithm steps:

1. Sort by keyword length, the longer the keyword, the higher the priority. Then the sorted key sequence K is obtained.

2. According to the rule of extracting the matched value of each keyword, the extraction rule set R is obtained, and the value obtained of the keyword A according to R is marked as VA.

3. Traverse the string S from the beginning to the end, and get the Levenshtein distance LiA at the i-th character with the keyword A, which ends with i-th character.

4. For fault tolerance, set the threshold for the Levenshtein distance. Regard the candidate word matches keyword A when the distance between the candidate word and keyword A is less than the threshold. When both LiA and LiB are smaller than the threshold, the priority of A and B is compared in K, and the keyword with higher priority is taken.

5. The end positions of the n matched keywords in 4 divides the string S into n+1 segments, and the j-th keyword A matches the rule R in the j+1th segment of the string to find VA.

3. Experimental Analysis

3.1. Experimental Results of Image Processing

The image with red stamp interference as shown in Fig. 1 may encounter scrambling error in character recognition. Therefore, HSV model is used to describe the color image first, then the red region and other background clutter are extracted and subtracted from the original image to remove the interference information of the stamp. The effect is shown in Fig. 4.

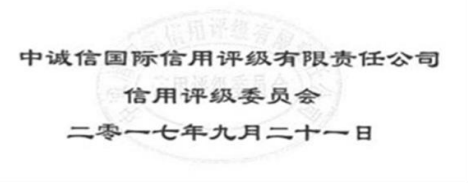

Fig. 4. Image fading processing.

After the red region is extracted, it is segmented according to the gray difference between the red region and the target, and the red region is completely removed. The final result is as shown in Fig. 5. The image shown in Fig. 5 is classified by character recognition, and no random code appears.

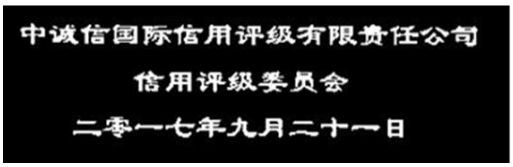

Fig. 5. Image difference calculation result.

3.2. Experimental Results of Character Recognition

For text recognition of bond information, the OCR model cannot be limited to specific keywords recognition because the bond information involves the contents of the company or organization name which can be almost any word. In order to verify the effectiveness of the proposed model, the text recognition experiment is carried out. In the experiment, the proposed model is compared with some existing models.

Dataset:

The amount of manually labeled data is not always adequate. In order to effectively train the model, the source of the data set contains two part. One is the collected and manually labeled bond manuscripts images; the other one is the automatically generated data, converting the collected text into an image, and the original text as labels. The amount of word in the data set is 500,000, and the test set contains about 50,000 words of bond manuscript.

Model:

For the OCR model in part B, the specific implementation parameters in this experiment are shown in Table 3-1. In the model, the activation function of each layer adopts ReLU, and the training adopts stochastic gradient descent (SGD). In order to speed up the training, batch normalization is added after the third and fourth convolutional layers.

Table 1. Optical Character Recognition Model Parameters

Type	**Parameters**
LSTM	256 Units
LSTM	256 Units
Convolution	Kernel：512，size：1x2，stride：1
Convolution	Kernel：512，size：3x3，stride：1
MaxPooling	Window Size：1x2，stride：2
Convolution	Kernel：256，size：3x3，stride：1
MaxPooling	Window Size：2x2，stride：2
Convolution	Kernel：128，size：3x3，stride：1
MaxPooling	Window Size：2x2，stride：2
Convolution	Kernel：64，size：3x3，stride：1
Input	Height:32

On the collected data, the experiment is conducted. The experimental results are shown in Fig. 6. In the experimental results, it is shown that the proposed model performs best.

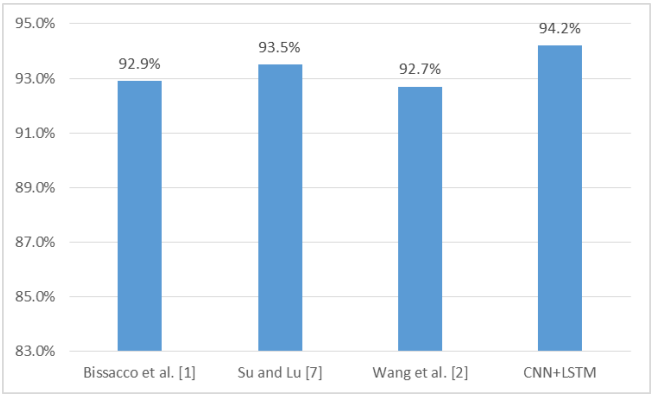

Fig. 6. Recognition accuracy.

3.3. Fault-Tolerant Information Extraction Experiment Results

As can be seen from the experimental results of 3.2, there is more or less errors in the OCR recognition result. Therefore, extracting information in such text has a certain probability that key information cannot be correctly extracted. According to statistics, 67% of the recognition errors are the key words with only one character wrong. Therefore, we set the tolerance threshold to 1 and use the key information extraction method in 2.3 to extract as much information as possible. Based on the OCR recognition of 3.2, the key information is extracted, and the accuracy is 96.9%.

4. Conclusion

There are two main ways to store manuscript information, one is text form, the other is image format. For text format manuscripts, the text information can be extracted directly, and the key-value model can be established by natural language processing. For image format manuscripts, OCR character recognition is needed. Aiming at various kinds of disturbances in the manuscript images, the difference subtraction operation of HSV color model is used to remove the disturbance information in the images. The recognition rate of the characters is as high as 94% by using the graphic sequence model. It has high practicability. No matter what kind of storage form of manuscript information needs to be processed by natural language to extract the correct keywords and their key values. This paper uses Levenshtein distance algorithm to extract the keywords. Its accuracy is as high as 96%. In a word, the recognition rate of the key information of manuscript is as high as 90%.

References

[1] Bissacco, A., Cummins, M., Netzer, Y., & Neven, H. (2013). Photoocr: Reading text in uncontrolled conditions. *ICCV.*

[2] Wang, T., Wu, D. J., Coates, A., & Ng, A. Y. (2012). End-to-end text recognition with convolutional neural networks. *ICPR.*

[3] Jaderberg, M., Simonyan, K., Vedaldi, A., & Zisserman, A. (2015). Reading text in the wild with convolutional neural networks. *IJCV.*

[4] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *NIPS.*

[5] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradientbased learning applied to document

recognition. *Proceeding of the IEEE, 86(11)* (pp. 2278-2324).

[6] Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H., & Schmidhuber, J. (2009). A novel connectionist system for unconstrained handwriting recognition. *PAMI, 31(5),* 855-868.

[7] Su, B., & Lu, S. (2014). Accurate scene text recognition based on recurrent neural network. *ACCV.*

[8] Bengio, Y., Simard, P. Y., & Frasconi, P. (1994). Learning longterm dependencies with gradient descent is difficult. *NN, 5(2),* 157-166.

[9] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9(8),* 1735-1780.

[10] Gers, F. A., Schraudolph, N. N., & Schmidhuber, J. (2002). Learning precise timing with LSTM recurrent networks. *JMLR, 3,* 115-143.

[11] Eam, K. T., & Dinesh, P. M. (1995). A transputer-based automated visual inspection system for electronic devices and PCBs. *Optics and Lasers in Engineering, 12(8),* 161-180.

[12] Francesco, A., Filippo, A., & Attilio, D. N. (2009). An online defects inspection system for satin glass based on machine vision. *Proceedings of International Instrumentation and Measurement Technology Conference.*

[13] Boult, T. E., Gao, X., & Micheals, R. (2004). Omni—Directional visual surveillance. *Image and Vision Computing, 22(7),* 515-534.

Jizhe Dai comes from Shanghai, China, born in 1993 Jiangsu China, graduated from University of Shanghai for science and technology and serve on the CFETS, major on image process.

Zhengyan Ma comes from Shanghai China, born in 1993 Anhui China graduated from Shanghai University and serve on the CFETS, major on natural language processing.

.