2022 IEEE International Conference on Visual Communications and Image Processing (VCIP) | 978-1-6654-7592-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/VCIP56404.2022.10008832

ACCR: Auto-labeling for Ancient Chinese Handwritten Characters Recognition on CNN

1st Peikun Wu
School of Mathematical Sciences
Nankai University
Tianjin, China
peikun.wu@foxmail.com

2nd Xin Yang
School of Mathematical Sciences
Nankai University
Tianjin, China
yangx1610114@163.com

3rd Fuhao Guo
School of Mathematical Sciences
Nankai University
Tianjin, China
2290256471@qq.com

4th Li Wang
China Institute of Science and Technology Information
Beijing, China
wl@istic.ac.cn

5th Cihan Ruan
Department of Computer Science Engineering
Santa Clara University
Santa Clara, U.S
cruan@scu.edu

***Abstract*—Chinese Character Recognition(CCR) is a critical application of Optical Character Recognition(OCR), a vital area of pattern recognition. Research on CCR in the past decades mainly focused on the modern Chinese characters, but not on the ancient ones. Compared to modern Chinese characters, ancient characters are more diverse and multiple ancient characters can correspond to one modern character. When doing recognition, the unique features of ancient Chinese characters cause a significant amount of time on manual labeling. This paper proposes an automatic labeling algorithm based on a semi-supervised dictionary training neural network that drastically decreases human effort. We first created an offline training set as a dictionary including 8,226 Chinese characters from ancient documents in modern fonts. And put the set into the network. Then we recursively retrained the network on an unlabeled data set of about 1.3 million characters images segmented from ancient documents resulting in a very high accuracy rate of 98.96%. This work is one part of our wide recognition of ancient documents with handwritten Chinese characters project.**

***Index Terms*—ancient documents, Chinese character recognition, handwritten recognition, residual network (ResNet), transfer learning**

I. Introduction

As the most used language in the world, Chinese is different from other languages. Chinese characters are neither alphabetic nor syllabary. They are logograms, which means almost every character has (or had) an independent meaning.

Though the actual number is uncertain, the number of unique Chinese characters employed throughout history is believed to be in the millions. The *Dictionary of Chinese Character Variants* published by the Taiwan Ministry of Education in 2004 has 106,230 Chinese characters, which is the highest record in a Chinese dictionary. However, Chinese characters have kept on evolving over the past thousands of years, from the Zhou Dynasty (221-106 B.C.) to present. The revolution includes not only the writing materials but also the fonts. More often than not, the characters go through the simplification of strokes. For example, the Chinese character "Dragon" has 17 strokes in ancient Chinese and later evolved into the modern Chinese character we know today with only five strokes (Fig. 1). Only a tiny percentage of Chinese characters are still in use today. According to the Chinese official institute, around 3,500 characters are used in daily life. Many modern characters can be traced back to their pictographic ancestors over two thousand years ago.

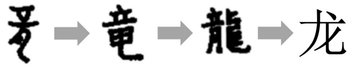

Figure 1. The revolution of Chinese character "*Dragon*"

A vast amount of ancient Chinese documents are preserved to the modern age. To read these ancient books, Chinese character recognition is a critical part. Chinese character recognition(CCR) first start in the year 1966(1), and researchers have already developed mature technologies around it. Multiple techniques, i.e., hierarchical template matching, nonlinear normalization, decision tree classification, and multiple classifier fusion, have been applied to such topics in the past, and researchers have achieved impressive results(2). In recent years, with the development of deep learning, neural networks have shown some remarkable achievements in OCR(3).

Compared to modern Chinese, ancient Chinese character recognition is a much harder field due to the variety and huge amount of data. A neural network is a powerful tool because of its ability to handle different shapes of characters and a large amount of data. To get a wide recognition of ancient documents with handwritten Chinese characters, a robust and accurate training set that is effective in improving test accuracy is necessary(4). However, to label a training set manually is an almost impossible mission. Xu et al. (5) established a database named CASIA-AHCDB, consisting of a large scale of ancient Chinese characters. During the database creation

978-1-6654-7592-1/22/$31.00 ©2022 IEEE

process, they used the annotation mechanism based on the Bayesian framework and geometric models proposed by Yin et al.(6) to give the matching score of the testing data and corresponding corpus. Although the accuracy of this method is not bad, the complexity is equally high. Moreover, for a traditional method, it is difficult to find a robustness feature to adequately describe characters, while deep learning can do it quickly.

Our work proposed a new algorithm to automatically generate a large capacity annotated training set based on a convolutional neural network. It is a semi-supervised scheme based on the transfer learning of ResNet, but we simplified the process of building a training set with annotation.

The remaining content is organized as follows. Section II reviews our pre-processing works on auto-cropping for our original dataset. Section III describes the details of the proposed auto-label algorithm. Section IV presents our experimental results and analysis. Last but not least, we give the conclusions in Section V.

II. Pre-processing

To apply a semi-supervised neural network to the automatic labeling scheme, a labeled training set and a larger unlabeled one are both required. In this section, we present the preparation of the study on how to generate those two data sets.

A. Process of labeled data set creation

We analyzed 8,226 most frequently occurring Chinese characters from some ancient documents that have been digitized and numbered. Then for each character, we used 18 existing kinds of fonts, including simplified and traditional versions, to generate 18 64×64 pixels image for each font, which sums to 149,068 images to form the labeled training set S_0 as shown in Fig. 2. Furthermore, to better simulate the real ancient documents, we added random noise and distortion to the images.

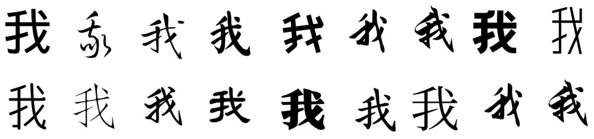

Figure 2. An example of the labeled data for Chinese character "*I/me*"

This is an easy and effective method considering the inheritance from ancient to modern Chinese characters. We used different shapes of one character as the initial training of the neural network to ensure that the algorithm could perform better at extracting the features.

B. Process of unlabeled data set creation

We selected two types of well-preserved ancient documents: poetry (e.g. *Shanyang Yi Gao*) and history books (e.g. *Liang Shu*,*Sui Shu* from *Emperor's Four Treasuries*, the largest official collection of Chinese historical books).

To minimize the impact of data quality on the final recognition result, we specially developed an image segmentation algorithm based on a hybrid of histogram projection and Gaussian heat map. Fig. 3 shows the steps. It is more flexible and efficient for document fragmentation. We named the algorithm as *AHAmap* (An Adaptive Chinese Handwritten Character Segmentation Algorithm Based on Histogram Projection and Gaussian Kernel Convolution Map), and the technical details were described in a related paper.

Characters' pictures from the selected ancient documents were unified into 64×64 pixel-sized images, which make up the unlabeled data set T.

C. ResNet

Since the convolutional neural network (CNN) can efficiently extract features, it is widely used in pattern recognition nowadays. Plus, the residual neural network (ResNet) deals with gradient descent, making it possible to increase the number of layers of CNN networks.

According to Liu et al.(7), ResNet-50 achieves good results on modern Chinese handwritten text recognition with 90% accuracy, which is why we chose transfer learning. In fact, in the process of transfer learning, we shrunk down the layers of ResNet-50 by half and obtained a higher accuracy rate, which will be discussed later in the article.

III. The Proposed Method

Since we already had a labeled training data set S_0, and an unlabeled one T, we set up the semi-supervised network as dictionary learning on ResNet(Fig. 4.).

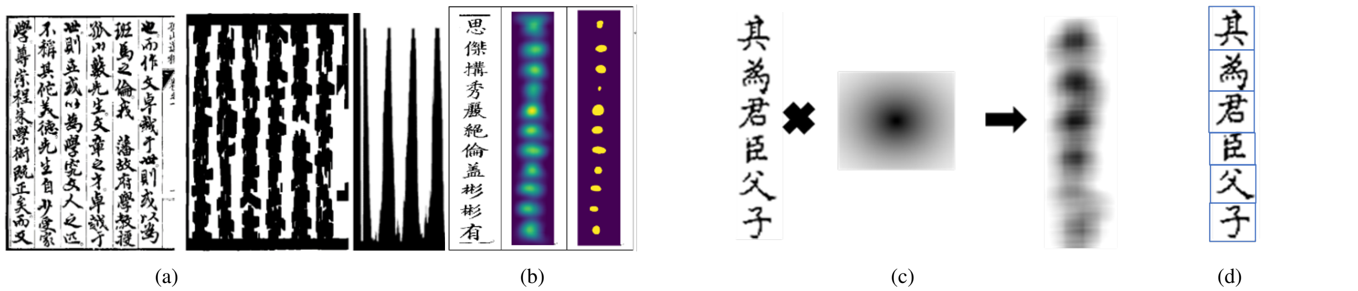

Figure 3. Unlabeled data set creating process (a) Binarization and Histogram. (b) Generate a Gaussian heat map. (c) Gaussian probability map convolution for text line segmentation. (d) Character segmentation.

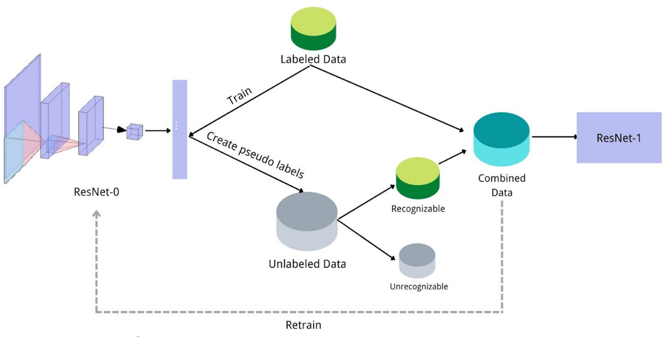

Figure 4. Design of this proposed semi-supervised dictionary learning scheme.

A. Process of dictionary learning

As a semi-supervised self-training network, we employed ResNet on our initial labeled data set S_0 as the process of dictionary learning. ReLU was used as the activation function, and the loss function we used as Baidu's warp-ctc(8) which is described as Connectionist Temporal Classification (CTC) loss function (9). We denoted the trained network as ResNet-0. The flowchart of the ResNet for ACCR is shown in Fig. 5.

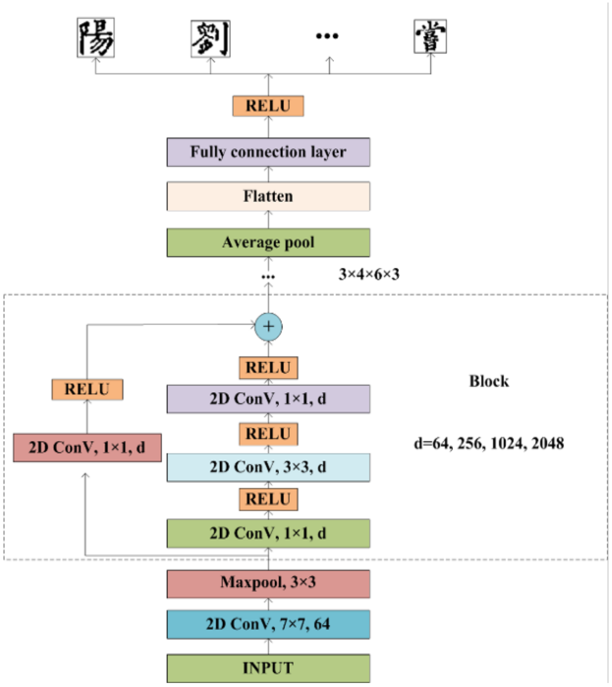

Figure 5. The structure of the ResNet for the proposed algorithm.

After that, we applied the ResNet-0 on the unlabeled data set T to generate pseudo labels. Then those data would be divided into two parts: recognizable T_1 and unrecognizable $\hat{T}_1$. Later, we combined T_1 and the dictionary S_0 to reach a larger labeled data set S_2 and retrained the ResNet-0 with the new data set. The retraining stopped when the accuracy reached 99% to avoid, overfitting and a better network named ResNet-1 was generated.

Finally, we employed this ResNet-1 on the $\hat{T}_1$ and found that the recognition ability of the new model was greatly improved.

B. Pseudo code

```
Details of ACCR algorithm
Input The initial labeled dictionary set $S_0$ and normalized unlabeled set $T$.
Output A labeled data set $S_k$.

Step1 Train ResNet-0 based on $S_0$ and apply it on $T$ to obtain two subsets: recognizable and unrecognizable, named as $S_0$ and $\hat{T}_1$.
    1.1 Transfer learning ResNet-50, train it on dictionary $S_0$ and treat it as ResNet-0.
    1.2 Score the similarity between $X_k \in T$ and $y_i \in S_0$ by likelihood probability classifier.
    1.3 Set a threshold $P$ for the classifier, to check if $p_{i_0}(X_k) = max(p_i(X_k)) > P$, put the $X_k \in T$ into recognizable set $T_1$ or $\hat{T}_1$.
    1.4 Check the category results.
    1.5 Evaluate $T_i$ after inspection and measure by using $|T_1| \setminus |T|$.
        If $|T_1| \setminus |T| \geq 0.05$, then merge $T_1$ and $S_0$ into a new data set $S_1$ as a updated dictionary for ResNet-0.
        Else Decrease the threshold $P$ to avoid overfitting or quit.
Step2 Apply $S_1$ and $\hat{T}_1$ in Step1 and receive ResNet-1, $S_2$ and $\hat{T}_2$ iteratively.
```

We recursively repeated the identification of the unlabeled dataset and merged the recognized ones into the dictionary until the group of unlabeled characters was almost empty or no new characters were recognized. As we trained ResNet repeatedly using an ever-growing dictionary, the accuracy rate kept increasing, as shown in Table I.

Table I
RECOGNITION ACCURACY OF TRAINED RESNET-0 AND RESNET-1 ON A SAMPLING UNLABELED SET T'

	Training Accuracy	Recognition Accuracy Rate
ResNet-0	99.96%	56.00%
ResNet-1	99.98%	64.22%

Moreover, We can improve the accuracy rate by adjusting the P value. The initial value of P was set to 0.7 and then reduced to 0.5, which eliminated the manual correction step and reduced the labor cost. To avoid overfitting, we controlled the accuracy rate during the dictionary learning.

IV. EXPERIMENT RESULTS

To verify the performance of semi-supervised dictionary training on auto-labeling of ancient Chinese characters, we employed this algorithm on our dictionary: a labeled data set S_0 including 149,068 modern Chinese characters both in the

Table II
RESULTS OF ACCR'S TRAINING PROCESSES

Batch No.	Input Characters	Qualified Chracters	Recognizable $\lvert T_i \rvert$	Unrecognizable $\left\lvert \hat{T}_i \right\rvert$	Unlabeled Set $\lvert T \rvert$	Recognition Rate $\lvert T_i \rvert \setminus \lvert T \rvert$ %
1st	210,066	160,615	96,273	64,342	64,342	59.94
2nd	175,362	125,435	137,494	52,283	189,777	72.45
3rd	343,785	264,840	244,216	72,907	317,123	77.01
4th	510,976	446,219	412,163	106,999	519,126	79.40
5th	65,888	50,384	137,411	19,972	157,383	87.31
6th	154,428	142,287	149,862	12,397	162,259	92.36
Total	1,460,505	1,189,780	1,177,419	328,900	–	98.96

simplified and traditional format, and an unlabeled data set T including 160,615 ancient Chinese characters cropped from the selected ancient documents. The data in both sets are all 64×64 pixel-sized images.

Experiments were run on a GeForce GTX TITAN 12G GPU, which has a Linus 64-bit operation system. There are 4 kernels of Intel(R) Xeon(R) Platinum 8269CY CPU @ 2.50GHz 3.10 GHz with a 16 GB memory available.

We initially set the $P = 0.7$. As a result, 96,273 from T were successfully recognized and delivered to the corresponding data set T_1, while the rest was sent to $\hat{T}_1$. As a preliminary stage of the experiment, we used manual verification to measure the effects of deep learning.

We found that, for the first training batch, the size of recognizable data set is $|T_1| = 96273$, and unrecognizable set is $\left|\hat{T}_1\right| = 64342$. We used the formula $|T_i| \setminus |T|\,\%$ to calculate the recognition rate, and the result of the first batch is 59.94%.

After the training of the first batch, we received the ResNet-1, and we put the recognizable set $T1$ into the dictionary set and got S_1. The unrecognizable set $\hat{T}_1$ was merged with 125,435 more new unlabeled data from the second batch. The new unlabeled data set was created.

We partitioned all the characters' segmentation into six batches according to the age of the original ancient documents and added them to the training by repeating the steps above iteratively. The older the document, the greater differences between the character forms contained in it and modern forms may be. Each batch represents an increase in the difficulty of identification in general. It needs to be clarified that data that are corp incorrectly will be sorted directly into an unrecognizable set. In that case, the quality of image segmentation will decrease the recognition effect. In our experiments, there was a significant drop in recognition rate, which is caused by inappropriate segmentation. After six cycles of training, we found that there is a large ratio of inappropriate segmentation exist in the remaining unlabeled data set, so we terminated the experiment. As a result, the total amount of correctly recognized characters has reached 1,157,419, which is 98.96% of the qualified character images' segmentation. Because there are almost no relevant neural networks targeting ancient handwritten Chinese character recognition, when directly comparing the results of other networks originally focused on modern Chinese recognition, our test results have a very significant advantage. The results are shown in Table II.

We believe that with the improvement of character segmentation quality, our total recognition accuracy will increase. Also, because of the features of the neural network, we have the flexibility to expand the size of the unlabeled data set to generate a large-scale annotated ancient handwritten Chinese character database. This does our next work on the CCR of ancient documents possible.

V. CONCLUSION

In this paper, we innovated an auto-labeling algorithm for ancient written Chinese characters based on ResNet that applied the semi-supervised dictionary learning structure. To reduce labor costs and improve accuracy, we innovatively created a labeled character set by using Chinese fonts that are encoded in Unicode format. Because of the evolution and association between ancient and modern Chinese characters, this data generation process is more efficient and simpler than the process used in previous works. Next, we built another unlabeled data set for the network from selected ancient Chinese documents. We used a self-developed character segmentation algorithm to generate the data set. After that, we applied both data sets on the ResNet and iteratively trained the network to recognize and label ancient Chinese written characters from learning modern ones. There was a significant increase in recognition rate as the number of training sessions increased. According to the results of experiments, the algorithm proved to be realistic and feasible. It also provides a good basis for our further work on the handwritten Chinese character recognition project on ancient documents.

ACKNOWLEDGEMENT

The authors are grateful to Changyun Wang, Dr. Hongqi Han, Dr. Jun Xu, and Dr. Jishou Ruan for their contributions and Jerry Wu's help in this paper. Also, they appreciate the anonymous reviewers' insightful criticism and recommendations for raising the paper's excellence. This research was funded by the project of "OCR image detection and recognition of ancient characters based on deep learning mode" of the Institute of Science and Technical Information of China, Grant No. ZD2020/09-03.

References

[1] R. Casey and G. Nagy, "Recognition of printed chinese characters," *IEEE Transactions on Electronic Computers*, no. 1, pp. 91–101, 1966.

[2] R. Dai, C. Liu, and B. Xiao, "Chinese character recognition: history, status and prospects," *Frontiers of Computer Science in China*, vol. 1, no. 2, pp. 126–136, 2007.

[3] S. R. Narang, M. K. Jindal, and M. Kumar, "Ancient text recognition: a review," *Artificial Intelligence Review*, vol. 53, no. 8, pp. 5517–5558, 2020.

[4] C. Luo, X. Li, L. Wang, J. He, D. Li, and J. Zhou, "How does the data set affect cnn-based image classification performance?" in *2018 5th International Conference on Systems and Informatics (ICSAI)*, 2018, pp. 361–366.

[5] Y. Xu, F. Yin, D.-H. Wang, X.-Y. Zhang, Z. Zhang, and C.-L. Liu, "Casia-ahcdb: A large-scale chinese ancient handwritten characters database," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019, pp. 793–798.

[6] F. Yin, Q.-F. Wang, and C.-L. Liu, "Transcript mapping for handwritten chinese documents by integrating character recognition model and geometric context," *Pattern Recognition*, vol. 46, no. 10, pp. 2807–2818, 2013.

[7] C.-L. Liu, F. Yin, D.-H. Wang, and Q.-F. Wang, "Casia online and offline chinese handwriting databases," in *2011 International Conference on Document Analysis and Recognition*. IEEE, 2011, pp. 37–41.

[8] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*. PMLR, 2016, pp. 173–182.

[9] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[10] E. Mode. How chinese characters evolved. Youtube. [Online]. Available: https://www.youtube.com/watch?v=GpUqqtE2qUo

[11] Y. Li and Y. Li, "Design and implementation of handwritten chinese character recognition method based on cnn and tensorflow," in *2021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*. IEEE, 2021, pp. 878–882.

[12] F. Yin, Q.-F. Wang, and C.-L. Liu, "A tool for ground-truthing text lines and characters in off-line handwritten chinese documents," in *2009 10th International Conference on Document Analysis and Recognition*. IEEE, 2009, pp. 951–955.

[13] T. T. Ngo, H. T. Nguyen, N. T. Ly, and M. Nakagawa, "Recurrent neural network transducer for japanese and chinese offline handwritten text recognition," in *International Conference on Document Analysis and Recognition*. Springer, 2021, pp. 364–376.

[14] M. Yanchun, L. Yongjian, X. Qing, X. Shengwu, and T. Lingli, "Review of automatic image annotation technology," *Journal of Computer Research and Development*, vol. 57, no. 11, p. 2348, 2020.