Toward Automatic Recognition of Cursive Chinese Calligraphy

An Open Dataset For Cursive Chinese Calligraphy Text*

Jung Liang Wen-Hung Liao[1] Yi-Chieh Wu
Department of Computer Science
National Chengchi University
Taipei, TAIWAN
[1]email: whliao@nccu.edu.tw

Abstract—**Calligraphy is one of the most important writing tools as well as cultural heritage in ancient China. Compared with other calligraphy styles, the cursive script is least restricted and oftentimes exhibits the personality of calligraphers. However, this style-oriented expression makes the cursive script hard to recognize even for trained experts. The call for auxiliary tools for cursive Chinese calligraphy text recognition has thus arisen.**

Data play a key role in the era of deep learning, yet there is a lack of open databases for the cursive Chinese calligraphy. In this paper, we address this discrepancy by collecting 43000 images consisting of 5301 different cursive Chinese calligraphy text. We have augmented the database with basic image processing operations to obtain a training set containing a total of 656K images. After experimenting with several deep neural architectures, we provided a baseline model Enhanced M6 (EM6) as a proof-of-concept to tackle the classification task. The proposed EM6 model achieved 60.3% top-1 accuracy and 80.8% top-5 accuracy on the evaluation data set, an indication that deep neural network has the potential to undertake the mission of cursive calligraphy recognition.

Index Terms—**Cursive Chinese Calligraphy, Text Recognition, Deep Learning**

I. Introduction

Calligraphy is the traditional way of writing in Chinese culture, as well as a crucial part of Chinese art. Chinese calligraphy can roughly be categorized into five writing styles: regular script, clerical script, semi-cursive script, cursive script, and seal script. Among these styles, the cursive script simplified the text structures tremendously. Moreover, it relates most closely to the calligraphers, demonstrating more personal traits and philosophy. This leads to vast structure variations for the same character and makes it difficult for trained experts to recognize cursive writing effectively.

Deep learning-based approaches have achieved impressive results in handwritten Chinese character recognition [1] during ICDAR [2] contests, in which the dataset, CASIA HWDB [3] plays an important role. However, there exist no open datasets for cursive Chinese calligraphy text currently, which is the missing piece of the puzzle if we wish to apply deep learning to facilitate automatic recognition. In this work, we introduce the first-ever open dataset specifically for cursive Chinese calligraphy text, followed by several CNN-based models as a proof-of-concept to validate the feasibility of automating the recognition process. We have chosen the enhanced M6 (EM6) architecture as the baseline model after comparative performance analysis. We further evaluate the proposed EM6 model on 18668 cursive Chinese calligraphy images extracted from an external source: BiSouth model calligraphy and achieve 64.3% Top-1 accuracy and 80.5% Top-5 accuracy, respectively. Additional experiments have been carried out to ensure the robustness of EM6 under different test conditions.

The rest of this paper is organized as follows. In Section II, we review past datasets concerning handwritten text. We then dive into our dataset in Section III, including a brief introduction and potential issues. Afterwards, we investigated three different neural network architectures, namely, Enhanced M6 (EM6, derived from M6 [4]), DenseNet-18 [5], and 3-way neural network and compared their performance in Section IV. We then describe our experiment settings and present results in Section V. We make a brief conclusion and discuss future works in Section VI.

II. Related Work

In this section, we review past datasets related to handwritten characters as we consider cursive script as a variant of handwritten text.

A. MNIST database

The MNIST database [6] of handwritten digit images is one of the most widely used datasets in the field of text recognition and machine learning. Consists of handwritten digits 0 9 with a total number of 60000 images in training set and 10000 in the test set, the MNIST database is often considered as the hello world example for recognition task as well proof-of-concept for new approaches. Variants such as EMINIST [7] and Fashion-MNIST [8] have also been introduced to further increase the complexity and diversity.

*The cursive Chinese calligraphy (ccc) database is available at: https://github.com/nccuviplab/CursiveChineseCalligraphyDataset

978-1-7281-5453-4/20/$31.00 2020 IEEE

B. CASIA HWDB 1.1

In the field of handwritten Chinese text, CASIA HWDB [3] is one of the most indicative datasets. Built by the National Laboratory of Pattern Recognition (NLPR), Institute of Automation of Chinese Academy of Sciences (CASIA) in China, the dataset contains both the online and offline datasets. Since we are not able to extract the online trace or dynamics of writing from ancient calligraphy work, we focus on the offline part of the dataset, the CASIA HWDB 1.1.

The CASIA HWDB 1.1 consists of 3755 different classes with around 1170K images in total. It was collected through crowd sourcing by 300 writers and is also the designated dataset of ICDAR Chinese Handwriting Recognition Competition in 2011 and 2013 [2]. Samples from the HWDB 1.1 are shown in Fig. 1.

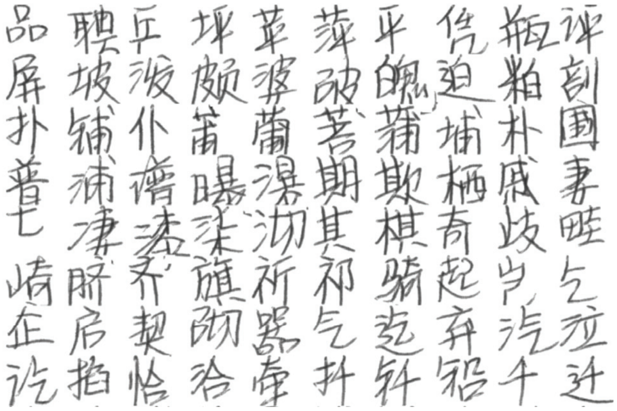

Fig. 1. Samples from CASIA HWDB 1.1

III. The Cursive Chinese Calligraphy Dataset

In this section, we give a brief overview of our dataset, and some issues that we have identified after further investigation.

A. Overview of the dataset

Our cursive Chinese calligraphy dataset consists of 43000 grayscale images of size 96x96 for 5301 different Chinese characters. As we further investigate the dataset, two major issues have come to the fore: 1) data imbalance and 2) variants of the same character that are structurally dissimilar.

B. Issues about the dataset

1) Data imbalance problem: As shown in Fig. 2, the number of images in each class varies from 1 to 70. Not only does each class contain insufficient samples, but there is also a severe data imbalance problem. One solution to this issue is to apply image augmentation on the training set so that each class contains similar amounts of images. We will discuss the image augmentation process in Section V.

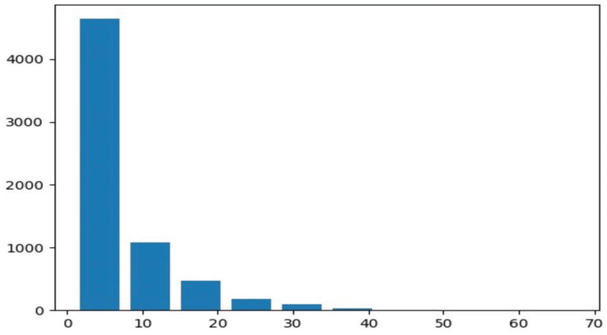

Fig. 2. Distribution of our proposed dataset. The x-axis indicates the number of images, where the y-axis indicates the number of classes corresponding to the number of images.

2) Variants of the same character: As depicted in Fig. 3, another issue that comes to our mind is that characters in the same class can appear very different structurally. This happens since the usage of characters could change in different dynasties. The structural difference makes the decision boundary of machine learning models hard to train. Therefore, we decide to refine each class so that variants will be split into a new class, as demonstrated in Fig. 4. This manual procedure is conducted by 5 people with background in Chinese calligraphy. The total number of classes after this process increased to 6494.

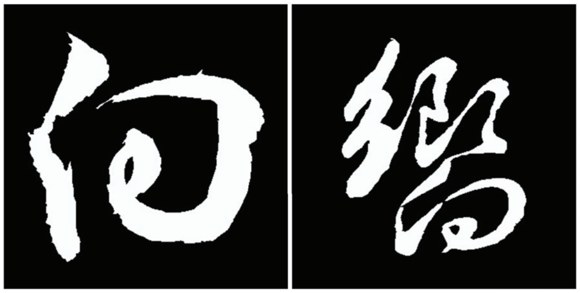

Fig. 3. Two distinct variants of the same character

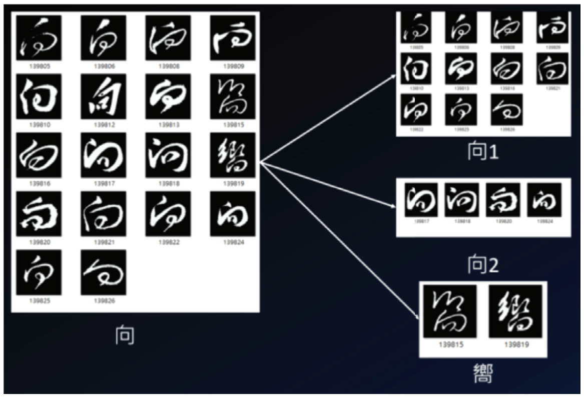

Fig. 4. Class refinement process

IV. DNN Models for Cursive Text Recognition

As there lacks prior research about cursive Chinese calligraphy recognition, we consider cursive script as a special case of handwritten characters. Although the MCDNN architecture proposed in [1] achieved a satisfactory result, the model might be too large for our dataset as there is an apparent gap of the number of images between our dataset and HWDB 1.1. As a result, we investigated several lighter-weight architectures: Enhanced M6, DenseNet-18 and the 3-way neural network. We will compare the performance in terms of accuracy in Section V.

A. Enhanced M6

The EM6 is constructed by adding batch normalization and an additional fully connected layer to the original M6 [4] architecture in to reduce the impact of over-fitting. An overview of our Enhanced M6 model is shown in Fig. 5.

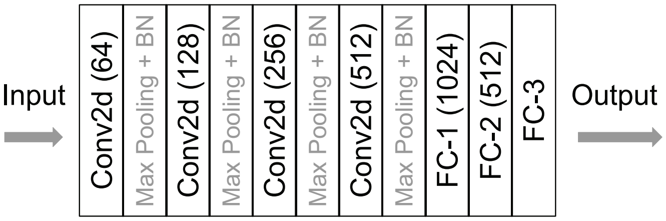

Fig. 5. Network architecture of Enhanced M6

B. DenseNet-18

DenseNet [5] applies the idea of dense connection, where connections among all layers are established in a feed-forward fashion. The main advantages of DenseNet are to alleviate the vanishing gradient problem as well as reducing the number of parameters. We simplified the original DenseNet-121 to an 18-layer-in-depth version, denoted as DenseNet-18, as the former was designed for ImageNet.

C. 3-way neural network

The 3-way neural network is devised based on our observation of Chinese writing, where one normally writes from top to bottom, and from left to right. The 3-way neural network consists of three feature extraction branches. First of all, the whole text image is fed to a convolutional network for global feature extraction. Then we split the image into Left and Right (as well as Top and Bottom, respectively), feed each slice into a shallow CNN block for feature extraction. We treat the features from the two branches as sequential data and feed them to a LSTM layer. Finally, the output of these branches are combined with the holistic features, and fed to the prediction layer. The architecture for the proposed 3-way network is given in Fig. 6 and 7.

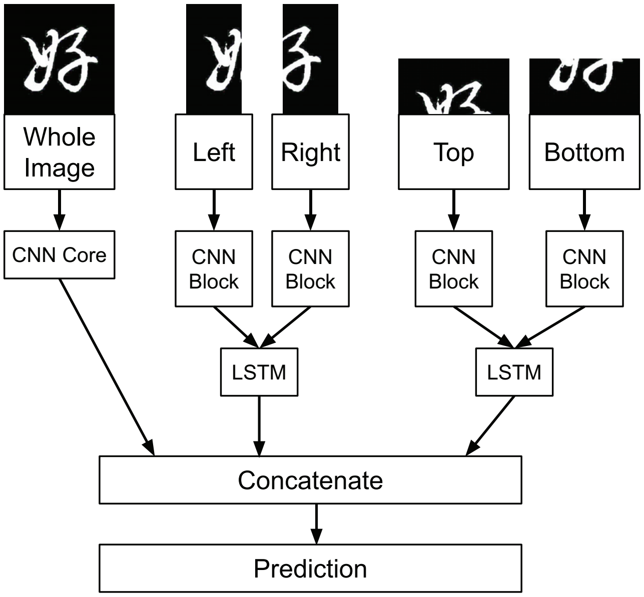

Fig. 6. The 3-way network architecture model

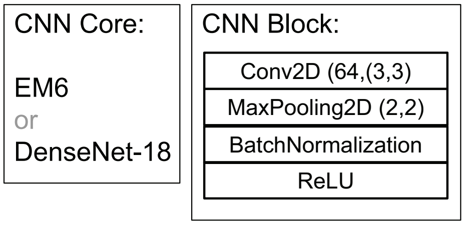

Fig. 7. Components of the 3-way network architecture

V. Experimental Settings and Results

In this section, we discuss the experiment settings regarding the recognition task. We then present and compare the performance of the three deep neural networks on our calligraphy database.

A. Training-test split

We first split the dataset to training and test subsets with an 8:2 ratio, where the classes contain less than 3 images remain only in training set. This leads to the consequence that the test subset contains only 9548 images from 3660 classes.

B. Data augmentation for the training set

As mentioned in the previous section, the total number of our dataset is insufficient to train the complicated deep neural networks. Moreover, the number of images in each class is severely imbalanced. Therefore, we apply the image augmentation technique on the training subset so that each class contains at least 100 images. The parameters of the augmentation process are random rotation within 15 degrees, vertical and horizontal shift within 5 pixels, and scale between 0.5 to 1.5. The training subset contains 656K images in total after the augmentation process.

C. Post-processing - the class aggregation process

Although we split variants out of its original class, the variants make no difference to the users. Therefore, for images derived from the same character, we merge their confidence afterwards and record the recognition result. We calculate the accuracy after this post-processing.

D. Evaluation criteria and results

The recognition task is evaluated based on Top-1 and Top-5 accuracy. The results including training and test accuracy are listed in Table I. The three networks demonstrated similar performance during the training phase. However, EM6 outperforms the others in the test phase, achieving 60.3% Top-1 accuracy and 80.8% Top-5 accuracy, and hence becomes our model of choice. As such, all the following evaluation will be performed using the EM6 architecture.

TABLE I
RECOGNITION RESULT DIFFERENT DNN MODELS

Model	Training Top-1 accuracy	Training Top-5 accuracy	Test Top-1 accuracy	Test Top-5 accuracy
EM6	96%	99.5%	**60.7%**	**80.3%**
3-Way Network (Core:EM6)	96.1%	99.4%	59.5%	79.1%
DenseNet-18	99.4%	99.9%	47.3%	67.5%
3-Way Network (Core:DenseNet-18)	94.0%	98.1%	59.7%	78.5%

E. Additional evaluation of the baseline model

As we have chosen the baseline model, we proceed to conduct further experiments to test its sensitivity about input size, training samples, as well as validating the generalization capability of the EM6 model. Specifically, we conducted experiments using input of three different sizes. We augment the training data with morphological operations including dilation and erosion and see how it affects the performance. Eventually, we validate our model's ability to generalize using the BISOUTH test data extracted from [9].

1) Changing input size: Theoretically, higher resolution input contains more information and features. In this experiment, we evaluate the effect of changing input sizes. The results are shown in Table II. Although image size of 96 × 96 yields best test accuracy, model performance using other resolutions seems comparable.

TABLE II
RECOGNITION RESULT OF EM6 MODEL USING INPUT OF DIFFERENT SIZES

Image Size	Training Top-1 accuracy	Training Top-5 accuracy	Test Top-1 accuracy	Test Top-5 accuracy
64 × 64	71.5%	89.7%	57.0%	78.4%
96 × 96	78.7%	93.7%	**60.3%**	**80.8%**
128 × 128	79.3%	94.0%	59.5%	79.6%

2) The effect of dilation and erosion during data augmentation: In a structural view of cursive Chinese calligraphy text, the structure of certain text might be break into many parts due to simplification. Therefore, we are curious if morphological operations such as dilation and erosion are able to recover some information regarding the original text.

In this experiment, we apply dilation and erosion to 5% of images during data augmentation process. Note that an image will only be subject to one type of operation, either dilation or erosion. The results using different augmentation sets are summarized in Table III. Slight improvement has been observed. Therefore, we further refine the baseline model with this new set of training samples to gain better accuracy.

TABLE III
RECOGNITION RESULT OF EM6 MODEL USING DIFFERENT AUGMENTATION METHODS

Augmentation Type	Training Top-1 accuracy	Training Top-5 accuracy	Test Top-1 accuracy	Test Top-5 accuracy
Shift, Rotate, Scale	78.7%	93.7%	60.3%	80.8%
Shift, Rotate, Scale, 5% Erosion, 5% Dilation	69.2%	87.3%	**61.0%**	**82.0%**

3) External data source evaluation – the BISOUTH test: In this experiment, we validate the generalization capability of the proposed EM6 model. We evaluate the model with cursive Chinese calligraphy text images from BISOUTH model calligraphy. A total of 18668 images of 2876 classes from 93 historical documents are extracted. We then evaluate our EM6 model on this test set. The results are summarized in Table IV, suggesting the robustness of the baseline model.

TABLE IV
RECOGNITION RESULT OF EM6 MODEL USING BISOUTH DATA

Training Top-1 accuracy	Training Top-5 accuracy	Test Top-1 accuracy	Test Top-5 accuracy
78.7%	93.7%	60.3%	80.8%

VI. CONCLUSION AND FUTURE WORK

In this work, we collected and processed the first open dataset for cursive Chinese calligraphy text. We then provide the Enhanced M6 model as a baseline for the recognition task after examining and comparing several deep neural network architectures. The proposed EM6 achieved 60.3% Top-1 accuracy and 80.8% Top-5 accuracy, leaving certain room for improvement for interested researchers.

Future work includes generating more training images using generative adversarial networks. (Preliminary results are illustrated Fig. 8.) Moreover, we need to investigate more modern neural network architectures and apply them to the recognition task.

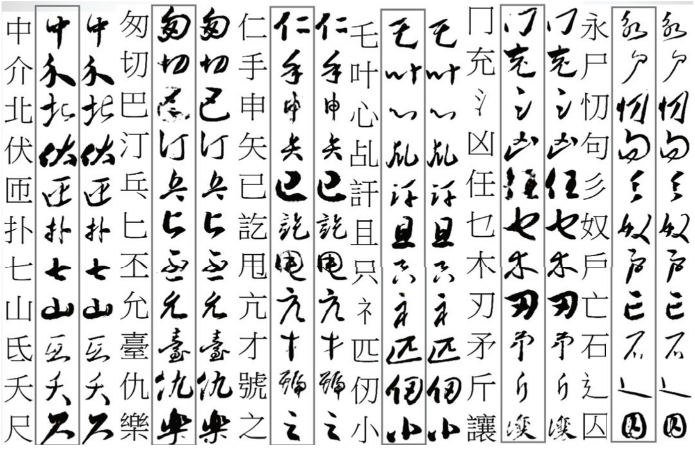

Fig. 8. Results of generating cursive Chinese characters (in gray box) using Pix2Pix [10]

Acknowledgment

We would like to extend our sincere appreciation to Mr. Chien Liu, the administrator of https://shufa.supfree.net/, for allowing us to scrape and re-distribute the hosted calligraphy images.

References

[1] Cirean, Dan, and Ueli Meier. ”Multi-column deep neural networks for offline handwritten Chinese character classification.” 2015 international joint conference on neural networks (IJCNN). IEEE, 2015.

[2] Yin, Fei, *et al.* ”ICDAR 2013 Chinese handwriting recognition competition.” 2013 12th International Conference on Document Analysis and Recognition. IEEE, 2013.

[3] Liu, Cheng-Lin, *et al.* ”CASIA online and offline Chinese handwriting databases.” 2011 International Conference on Document Analysis and Recognition. IEEE, 2011.

[4] Zhang, Yuhao. ”Deep convolutional network for handwritten chinese character recognition.” Computer Science Department, Stanford University (2015).

[5] Huang, Gao, *et al.* ”Densely connected convolutional networks.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[6] LeCun, Yann, *et al.* ”Gradient-based learning applied to document recognition.” Proceedings of the IEEE 86.11 (1998): 2278-2324.

[7] EMNIST: https://www.nist.gov/node/1298471/emnist-dataset

[8] Xiao, Han, Kashif Rasul, and Roland Vollgraf. ”Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms.” arXiv preprint arXiv:1708.07747 (2017).

[9] Bisouth model calligraphy: https://www.bisouth.com.tw/a-1.html

[10] Isola, Phillip, *et al.* ”Image-to-image translation with conditional adversarial networks.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.