Assessing Penmanship of Chinese Handwriting: A Deep Learning-based Approach

Zebo Xu[1], Prerit S. Mittal[2†], Mohd. Mohsin Ahmed[2†], Chandranath Adak[3*], Zhenguang G. Cai[1*]

[1]Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Hong Kong, SAR, China

[2]Department of Information Technology, Indian Institute of Information Technology, Lucknow, India

[3]Deptartment of Computer Science and Engineering, Indian Institute of Technology Patna, India

Author Notes

†P. S. Mittal and M. M. Ahmed contributed equally to this work.

This research was supported by a GRF grant (Project number: 14613722) from the Research Grants Committee of Hong Kong.

Upon acceptance of the paper, all resources will be available at https://github.com/datalabv01

*Correspondence should be directed to either Z.G. Cai, G05, Department of Linguistics and Modern Languages, Leung Kau Kui Building, The Chinese University of Hong Kong, Hong Kong SAR, email: zhenguangcai@cuhk.edu.hk, or to C. Adak, Department of Computer Science and Engineering, Indian Institute of Technology Patna, Bihta, Bihar - 801106, India, email: chandranath@iitp.ac.in.

Abstract

The rise of the digital era has led to a decline in handwriting as the primary mode of communication, resulting in negative effects on handwriting literacy, particularly in complex writing systems such as Chinese. The marginalization of handwriting has contributed to the deterioration of penmanship, defined as the ability to write aesthetically and legibly. Despite penmanship being widely acknowledged as a crucial factor in predicting language literacy, research on its evaluation remains limited and in the early stages, with existing assessments primarily dependent on expert subjective ratings. Recent initiatives have started to explore the application of convolutional neural networks (CNN) for automated penmanship assessment. In this study, we adopt a similar approach, developing a CNN-based automatic assessment system for penmanship in traditional Chinese handwriting. Utilizing an existing database of 39,207 accurately handwritten characters (penscripts) from 40 handwriters, we had three human raters evaluate each penscript's penmanship on a 10-point scale and calculated an average penmanship score. We trained a CNN on 90% of the penscripts and their corresponding penmanship scores. Upon testing the CNN model on the remaining 10% of penscripts, it achieved a remarkable performance (overall 9.82% normalized Mean Absolute Percentage Error) in predicting human penmanship scores, illustrating its potential for assessing handwriters' penmanship. To enhance accessibility, we developed a mobile application based on the CNN model, allowing users to conveniently evaluate their penmanship.

Keywords: Chinese handwriting, Penmanship, Deep learning, Convolutional Neural Network

INTRODUCTION

Handwriting in the digital age

Handwriting has been a cornerstone of human civilization; it enables people to record ideas and emotions through time. While some might think that modern technology has rendered handwriting archaic, it still plays a vital role in daily learning and communication; for instance, learning involves recording information using handwriting (e.g., note-taking). Students, especially young ones, predominantly rely on handwriting rather than digital typing for learning (Cutler & Graham, 2008; Jones & Hall, 2013; Sheffield, 1996). After all, children expand their vocabularies via reading and writing. Also, preschool learners acquire letters more effectively via handwriting compared to passive observation (James & Engelhardt, 2012), possibly because handwriting calls for a greater extent of sensorimotor involvement (Longcamp, Boucard, et al., 2006; Longcamp, Tanskanen, et al., 2006). University students were shown to engage in more active message recording if they take notes with handwriting than with typing (Mueller & Oppenheimer, 2014). With the availability of digital pens and handwriting input methods, modern digital technologies, contrary to common perception, have actually afforded handwriting an increasingly important role in information processing and communication.

However, the digital age has witnessed a detrimental decline in handwriting literacy (Jones & Hall, 2013; Medwell & Wray, 2008; Wollscheid et al., 2016). The prevalence of typing has particularly impacted the ability to handwrite Chinese characters (Tong & McBride-Chang, 2010), leading to challenging orthographic retrieval (Huang, Lin, et al., 2021; Huang, Zhou, et al., 2021), compromised fluencies, and crucially for this paper, a deteriorating penmanship. Penmanship refers to the ability to handwrite legibly and aesthetically (Fairbank, 2018). In particular, legibility pertains to the quality of being clear enough to read, while aesthetics refers to the beauty of handwriting style. Thus, in the context of Chinese handwriting, legibility

relates to the clarity of handwritten text (or, penscript) to ensure that the readers can easily understand the penscript without struggling to decipher it. For instance, regular script (or, 楷书 in Chinese) is a standardized and formal style of Chinese calligraphy (widely used in school education and daily communication due to its readability). This script features standardized stroke (or, 笔画 in Chinese) formations across characters (e.g., stroke form, length, direction, positions where the strokes are connected, and clear beginning/ending points for strokes). Different from legibility, the aesthetic aspect of handwriting focuses on the visual beauty derived from its style, strokes, and presentation, providing an artistic and visually appealing form of written expression. For example, the semi-cursive script (or 行书 in Chinese) diverges from regular script through its more varied stroke formations, facilitating greater room for artistic expression. This script is renowned for its partially connected strokes, which sometimes undergo simplification, combining or abbreviating specific strokes. Consequently, the script's more fluid and dynamic strokes allow for personal expression and interpretation, fostering the creation of its distinct aesthetic. These alternations offer a spectrum of aesthetic possibilities without (or slightly) compromising overall legibility (A more extreme example is the cursive script, celebrated for its aesthetic aspects but often lacking in legibility).

Under the age of the digital typing, the increased reliance on digital typing has been observed to impact the legibility of the penscript (Kiefer & Velay, 2016). Marquardt et al. (2016) conducted a large-scale survey among 1907 teachers from the primary and secondary schools, reporting that over 30% of girls and 50% of boys encountered difficulties acquiring legible handwriting. Among 1174 secondary school teachers, 93.2% of them reported that the biggest issue students faced in acquiring handwriting skills was the illegible penscript. Moreover, the appreciation for aesthetic qualities of handwriting has waned, contributing to a diminishing cultural and educational value placed on penmanship (Sang, 2023). In the context of Chinese handwriting, the loss of the aesthetic aspect is particularly concerning, as

penmanship (e.g., calligraphy) has long been celebrated as forms of artistic expression and Chinese cultural heritage (Wood, 1982). The decline in penmanship not only hampers individual handwriting skills but also erodes the cultural appreciation for the arts of Chinese character handwriting. While there have been some efforts to assess Indic penmanship (Adak et al., 2017), there are currently no assessments for evaluating Chinese penmanship. In this paper, we develop a deep-learning-based assessment system capable of generating human-like penmanship ratings for handwritten Chinese characters, offering a novel approach to preserving and revitalizing the aesthetic and arts of handwriting within the digital era.

Assessing handwriting

Handwriting entails the retrieval of orthographic codes and the stroke-by-stroke output of these codes as handwritten text. Consequently, there are multiple aspects of handwriting that can be assessed. One crucial aspect is the success or failure of orthographic retrieval (i.e., whether an individual retrieved the to-be-written character correctly from the orthographic long-term memory, independent of legibility and aesthetics). While this issue may not be considered significant in writing systems with substantial phonology-orthography correspondence (e.g., English), it is vital in writing systems without such correspondence. In Chinese and Japanese kanji, for example, individuals may struggle to retrieve the orthographic makeup of a character (a phenomenon known as character amnesia; e.g., Huang et al., 2021), making the assessment of orthographic retrieval success or failure essential for handwriting literacy.

Another aspect of handwriting that assessments have focused on is fluency. Skar et al. (2022), for instance, had test-takers (children) copy sentences as quickly and accurately as possible within a time limit, using the number of correctly written letters as a measure of handwriting fluency (similar assessments can be found in Berninger et al., 1991, and Peverly

et al., 2013). Some studies employed writing latencies to evaluate handwriting fluency. For instance, Rosenblum et al. (2006) demonstrated that copying latency between letters (duration from the offset of the last letter to the onset of the following letter) is a valid predictor for diagnosing children with handwriting difficulties. Martínez-García et al. (2021) found that the interval between the offset of the audio stimuli and the pen's first touch on the tablet was longer for children with handwriting difficulties than typically developing controls. Additionally, a recent study used letter writing duration and the number of pauses per letter to assess individuals' handwriting fluency in an alphabet writing task (Alamargot et al., 2020).

Although handwriting accuracy and fluency reflect the ability to retrieve orthographic representations and can thus serve as useful indices for handwriting literacy, they do not encapsulate penmanship, which concerns the quality of the end product of handwriting (i.e., penscript) (Fairbank, 2018). Importantly, there is evidence that penmanship contributes to language literacy development (Feder & Majnemer, 2007). It has been shown to be a significant factor in predicting individual spelling performance, reading ability, and mathematical ability (Eidlitz-Neufeld, 2003; Simner, 1988). There is an association between children's penmanship and their reading and spelling proficiencies (Caravolas et al., 2020). Furthermore, good penmanship might be an indicator of refine orthographic details, leading to better effects on language learning. Indeed, students with better penmanship tend to have better performance in their essay writing (Bull & Stevens, 1979). One possible reason for these associations could be due to good penmanship requires lots of handwriting practices, which is highly interactive with the linguistic processes (e.g., orthographic lexicon, semantic system, phonological lexicon and so on). For instance, previous studies on Chinese handwriting have been shown that lexical variables, such as character frequency and regularity etc, modulating the central and peripheral processes of handwriting (Wang et al., 2020; Huang, Zhou, et al., 2021; Huang, Lin, et al., 2021; neuroimaging studies: Yang et al., 2022; Li et al., 2023). Therefore, the end product of

handwriting: penmanship, may inherit crucial details of individuals' language competence, However, there are currently not many automatic tools for penmanship assessment.

Earlier penmanship assessments have employed subjective evaluations of penmanship, particularly for letter handwriting in alphabetic languages. In the Children's Handwriting Evaluation Scale (CHES; Phelps et al., 1985), children copy sentences by handwriting, and trained raters assess the legibility of the penscript on a 5-point scale based on four factors: letter forms, slant, space, and general appearance. Similarly, in the Minnesota Handwriting Assessment (MHA; Reisman, 1999), children copy English words, and occupational therapists evaluate the penscript on alignment, size, spacing, and form appearance (a similar approach was adopted in the Minnesota Handwriting Test; Reisman, 1993). In the Print Tool (Olsen & Knapton, 2006), test-takers handwrite uppercase letters in alphabetical order, dictated lowercase words, and numbers in words; occupational therapists then assess penmanship based on the orientation, placement, and size of handwritten letters in the penscript. There are, however, two potential problems with these subjective tools of penmanship assessment. First, these tests tend to be costly, requiring an expert to evaluate handwriting. Second, subjective penmanship evaluation is susceptible to individual biases. Therefore, an automatic penmanship assessment is still required to perform a more objective assessment.

Assessing handwriting in Chinese

Chinese uses a logographic writing system, with characters as the smallest free-standing meaning-carrying units. A Chinese character follows the "square script" (Chen & Kao, 2002), composed of one or more radicals in a specific spatial layout. For instance, the character 树 (*shu*$_4$, meaning "tree") consists of three radicals (i.e., 木, 又, and 寸) arranged horizontally, while the character 李 (*li*$_3$, meaning "plum") comprises two radicals (i.e., 木 and 子) placed vertically. Radicals are further comprised of strokes; for example, the radical 木 is made up of

一, 丨, 丿, and 丶. When learning to handwrite characters, children typically adhere to a strict order of radicals within a character and a stringent sequence of strokes within a radical.

Some assessments have been developed for handwriting fluency in Chinese. Chow et al. (2003) asked children to copy Chinese characters as quickly and neatly as possible within 5 minutes, measuring handwriting fluency by the number of characters written correctly per minute (Chow et al., 2003; Tseng & Hsueh, 1997). More recently, Li-Tsang et al. (2022) had children copy characters on a digital tablet and assessed handwriting fluency in terms of pen-on-paper time, pen-in-air time, and the number of characters per minute (see also Lam et al., 2011; Li-Tsang et al., 2022).

Assessments have also been developed for penmanship in Chinese handwriting. In Tseng's Handwriting Problem Checklist (Tseng, 1993), teachers evaluate the penmanship of children's handwritten characters from daily writing samples based on spacing, spatial relationships, size consistency, and radical/stroke appearance. Similarly, Chan et al. (2008) assessed penmanship using space between characters and character size variations. Some penmanship assessments examined how similar children's handwritings of their own names resemble typed characters (Chan & Louie, 1992; Tse et al., 2019).

Computerized assessment of penmanship

More recently, there have been attempts to use more objective computerized measures to predict penmanship. Rosenblum et al. (2004) studied children's writing on a digital tablet and showed that a longer pen-on-paper duration (as captured by the digital tablet) predicted lower handwriting legibility (as rated by professional evaluators on Hebrew handwriting; Erez & Parush, 1999). However, while demonstrating the correlation relationship between writing duration and legibility, this study did not delve into assessments of both aesthetic and legibility. Falk et al. (2011) had children copy words on a tablet (using MHA) and were able to identify

children with poor handwriting legibility utilizing the variation of letter height and distance of letter boundaries. Despite highlighting specific factors contributing to legibility, their approach does not offer a comprehensive computerized assessment of penmanship. Similarly, employing a random forest model, Asselborn et al. (2018) achieved a 96.6% accuracy in identifying children with poor handwriting legibility based on writing speed, pen pressure, inter-word gaps, and writing slant. These studies demonstrated that legibility can be predicted using computerized measures, However, they only focus on handwriting fluency and letter characteristics, while penmanship involved a much more complex cognitive process, especially in the context of handwriting in Chinese (e.g., the clarity of stroke forms, proper spacing between strokes/radicals, consistent sizing of strokes/radicals). Therefore, the indirect measures of handwriting (e.g., fluency and letter characteristics) in predicting legibility might not fully capture the handwriting penmanship.

A recent attempt to computerize penmanship assessment is Adak et al. (2017), who trained a convolutional neural network (CNN) to provide human-like penmanship assessment for Bengali penscripts. In particular, they were interested in testing whether their CNN model can perform like human beings in assessing aesthetics and legibility. They had participants given aesthetic and legibility scores respectively to the penscript documents; these data were imparted to the model and finally achieved an overall 86.01% F-Measure for the aesthetic analysis and 85.74% F-Measure for legibility prediction in the Bengali penscript documents. The same group also extended their work for understanding Bengali handwriting difficulties using handcrafted features with a Support Vector Machine (SVM) and auto-derived features with an inception network (Adak et al., 2018) and reinforcement learning (Adak et al., 2021).

In this paper, we take a similar approach to that of Adak et al. (2017) in developing a CNN for assessing the character penmanship (aesthetic and legibility) of traditional Chinese penscripts. Specifically, we leverage handwritten characters from an existing large-scale

traditional Chinese handwriting database. For ground-truthing, human raters provide a subjective penmanship rating for a character penscript. We trained a CNN model on a training set of character penscripts (along with their corresponding penmanship ratings). Subsequently, the CNN was assessed on a test set of unseen character penscripts, and its generated ratings were compared against human ratings for these unseen penscripts.

The Chinese handwriting database

The character penscripts (handwritten character images) utilized in this study were derived from a recent large-scale handwriting database (Cai et al., in prep), containing 48,000 penscripts of 1,200 characters from 40 native Cantonese-speaking college students (i.e., writers) in a dictation task. All the writers were undergraduate students at the Chinese university of Hong Kong. They were native speakers of Cantonese (mean age = 20.40 years, ranging from 18 to 27 years; 21 females and 19 males). These individuals resided in Hong Kong and predominantly used traditional Chinese characters and Cantonese as their dominant language from early childhood. All participants were typically developing writers, with none reporting difficulties in both reading and writing. Furthermore, the mean handwriting accuracy across all participants was 86.97%, ranging from 71.33% to 95.43%, indicating their general proficiency in handwriting. Additionally, all the participants were right-handed and had normal or corrected-to-normal vision and hearing. None of them reported having any neurological or psychiatric disorders/deficits, nor were they undergoing medical treatment. During the task, writers listened to a dictation phrase specifying the target character (e.g., "吩咐嘅咐," denoting the character "咐" in the word "吩咐"). They employed an inking digitizer pen (Wacom KP-130-00DB, Japan) to write down the target character on a paper sheet placed on top of an Intuos graphic tablet (Wacom PTH-651, Japan). Penscripts were captured as vector images using

OpenHandWrite (https://github.com/isolver/OpenHandWrite) and subsequently extracted for accuracy checking and penmanship rating.

Cai et al. (in prep) chose characters from the traditional Chinese Character Database (http://humanum.arts.cuhk.edu.hk/Lexis/lexi-can) and filtered the characters according to two criteria. First, they chose character candidates with an appearing frequency between 1 and 124414 according to the Database. They only considered characters with more than four strokes (leaving a candidate set of 2095 characters). For each of character, they identified their most common two-character word context and asked 21 university undergraduate participants (from the same universities as the participants in our study) to give word familiarity rating based on a scale of 1 to 7. They only included bicharacter words with an average familiarity equal or larger than 3.5. Dictation phrases were recorded using the text-to-speech reader Langdunv 7.6. The final choice of the characters was constrained by participant-hour considerations. In their study, they randomly included 1200 characters and each participant attended four sessions on different days, each session lasting for about one hour

METHODS

Below, we first report how we collected penmanship ratings for character penscripts, and then we propose a deep learning mechanism for automated penmanship assessment.

Penscript accuracy checking

We recruited three helpers (from the same population as the penmanship raters but not the raters themselves; see below) to check the accuracy of the penscripts (handwritten images) on the online survey platform Qualtrics (https://www.qualtrics.com). For each target character, the helpers were first shown the character in its typed format, followed by the 40 corresponding penscripts. They were instructed to code a penscript as "correct" if they recognized it as the

target character and "incorrect" if the penscript was not identified as the target character. Additionally, they were instructed to identify if a penscript, if correct, was an abandoned false start; if so identified, the false start strokes were later digitally removed from the image. We excluded all the incorrect penscripts, with a total of 39207 correct penscripts (with false starts removed) included in the penmanship rating (see below).

Penmanship rating

We recruited an additional 12 native Cantonese-speaking college students as penmanship raters (mean age = 21.25 years, ranging from 18 to 32 years; 8 females and 4 males). Raters first participated in a discussion session with two of the authors (Z.X. and Z.G.C.) regarding how to assess penmanship. The two authors initially selected 20 penscripts (of the same target character) that they perceived to vary in penmanship and used them as training items. The raters were asked to provide their penmanship rating on each of the 20 training penscripts on a 10-point Likert scale (1 meant to very poor penmanship and 10 meant very good) and then collectively discussed with the two authors to reach an agreement on the penmanship rating for each of the training penscripts. All raters concurred on the final ratings for the 20 training penscripts and were encouraged to use these as references in their subsequent penmanship ratings.

Following the training session, the 12 raters were randomly assigned to three groups of four raters. Each group rated all 39,207 correct penscripts (roughly evenly divided into 120 lists), resulting in each penscript receiving three penmanship ratings (i.e., one from a group). During the rating, a rater viewed a penscript image and provided a penmanship score on a 10-point scale.

To detect potential random responses by a rater, we incorporated two foil penscripts (one determined by the authors to exhibit poor penmanship and one to display good penmanship) in

each of the 120 lists (yielding a total of 240 foils). Comparing the poor-penmanship and good-penmanship foil penscripts in each list for each rater, we found that the good-penmanship foil consistently received a higher rating than the poor-penmanship foil. Consequently, we did not discard any ratings.

We also conducted an intraclass correlation analysis on the 10-point scale ratings to assess inter-rater reliability. The results revealed an intraclass correlation coefficient (ICC) of 0.377, with a 95% confidence interval for ICC population values ranging from 0.371 to 0.383. This finding indicates a fair degree of agreement among the ratings provided by multiple raters, though the moderation correlation may also suggest that penmanship assessment can be subject to individual biases. We further explored the inter-rater agreement with a 2-point deviation criterion from the mean. This analysis resulted in 74.5% agreement, suggesting a good agreement among raters within this specific deviation range.

Automated penmanship assessment

We now describe the treatment of the penmanship ratings and the development of our CNN for penmanship assessment of Chinese character handwriting.

Problem formulation: A penscript image was shown to the human rater as mentioned earlier, where they provided a penmanship score s in a discrete Likert scale (Likert, 1932), ranging from the lowest score s_L to the highest score s_H, i.e., $s \in \{s_L, s_L + 1, s_L + 2, \dots, s_H\}, s \in \mathbb{Z}$. For each character sample, multiple human raters provided their scores s_i; for $i = 1$ to n. We computed the *mean penmanship score* $p = \frac{1}{n} \sum_{i=1}^{n} s_i$ and tagged it to each penscript sample. In our case, $s_L = 1, s_H = 10, n = 3$. As a matter of fact, $p \in \mathbb{R}$ and p is in the range of $[s_L, s_H]$.

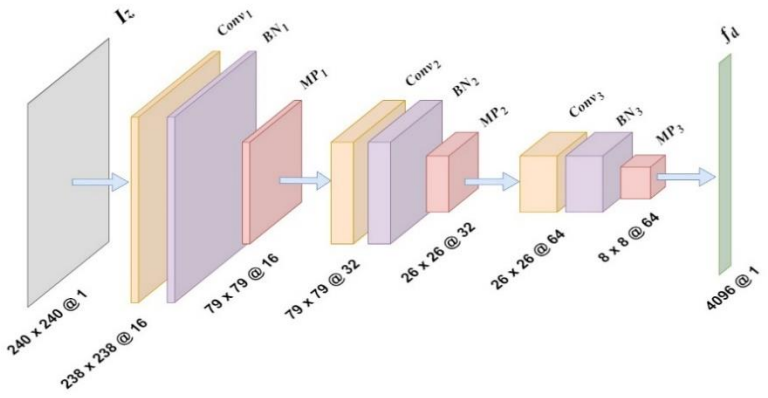
In our database, we have penscript images I_i and corresponding ground-truthed penmanship scores p_i; for $i = 1$ to m, where m is the total number of samples in our database.

To create a deep learning-based model that can mimic the human perception and provide a penmanship score of an unknown penscript, we formulate this penmanship score prediction task as a deep regression problem (Lathuilière et al., 2019), where we predict the penmanship score p of a given penscript image I.

Solution architecture: The given input image I was resized into I_z of size $n_z \times n_z$ without any loss of required information for our task while keeping the trace of the aspect ratio. This image size reduction helped to reduce training/computation time and to improve efficient memory management with limited hardware/GPU resources. For our task, empirically, we chose $n_z = 240$.

We employed a Convolutional Neural Network (CNN) for deep feature extraction, since it works better than traditional hand-crafted features for handwritten samples (Adak et al., 2019). The CNN-extracted features were fed to a Multi-Layer Perceptron (MLP) for the prediction of penmanship score (Goodfellow et al., 2016). The pictorial representation of our framework can be seen in Figure 1.

(a)



(b)

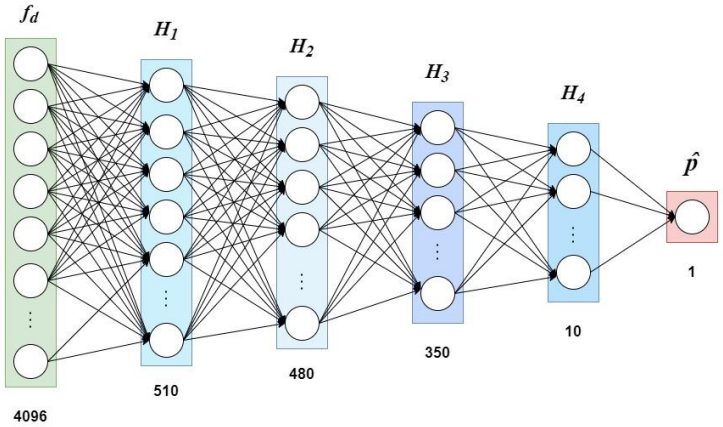


Figure 1. The proposed framework for (a) deep feature extraction, (b) penmanship score prediction.

CNN. The resized image I_z was input to the CNN that contains three convolution ($Conv_i$) layers, each followed by a batch normalization (BN_i) layer (Ioffe & Szegedy, 2015) and a max-pooling (MP_i) layer sequentially, for $i = 1, 2, 3$. In each $Conv$ layer, we used Rectified Linear Unit (ReLU) activation function to tackle complex non-linear patterns and to avoid the vanishing gradient problem (Goodfellow et al., 2016). In the $Conv_i$ layers, we chose the padding width (w_p^i), stride (d_s^i), number of filters (n_f^i), kernel size ($k_f^i \times k_f^i$), for $i = 1, 2, 3$. For the first $Conv$ layer (i.e., $Conv_1$), we chose $w_p^1 = 0$, $d_s^1 = 1$, $n_f^1 = 16$, $k_f^1 = 3$. For $Conv_2$ and $Conv_3$ layers, we fixed $w_p^2 = 2$, $d_s^2 = 1$, $n_f^2 = 32$, $k_f^2 = 5$ and $w_p^3 = 2$, $d_s^3 = 1$, $n_f^3 = 64$, $k_f^3 = 5$, respectively. We used BN layer, since it has regularization effect and is less bothered on weight initialization (Zhang et al., 2021). For all the MP layers, we employed a 3×3 sized kernel. We also used a dropout layer with 0.5 keep-probability after the third MP layer (i.e., MP_3) to reduce overfitting (Goodfellow et al., 2016). From the CNN, after MP_3 layer, we obtained 64 number of feature maps (Zhang et al., 2021) each of size 8×8, which we flattened and produced a 4096 ($= 64 \times 8 \times 8$) dimensional feature vector f_d.

$$f_d = \text{CNN}(I_z)$$

MLP. The f_d was embedded to an MLP with four hidden layers (H_i for $i = 1, 2, 3, 4$) added sequentially. The number of neurons in H_1, H_2, H_3, H_4 are 510, 480, 350, and 10 respectively, which were chosen empirically. In all the hidden layers, we used "He normal initializer" for weight initialization (Zhang et al., 2021). In H_1, H_2, H_3, H_4, we employed *L1*, *L2*, *L1_L2*, *L1_L2* regularizations to prevent the model from overfitting, where all regularization-parameters (λ) were tuned to 0.01 experimentally (Goodfellow et al., 2016). We also used two dropout layers, each with 0.5 keep-probability after H_1 and H_3. The output layer containing only one node is added sequentially to H_4, since we wanted to predict only one penmanship score for an input penscript image. In all the layers of MLP, ReLU activation function was used to attain non-linearity for handling arbitrary complex patterns. Here, the MLP worked as a deep regression model to predict the penmanship score $\hat{p}$.

$$\hat{p} = \text{MLP}(f_d)$$

Loss (L). To train the model, we employed Mean Squared Error (MSE) loss, which is defined as below.

$$L(\theta) = \text{MSE} = \frac{1}{m_{tr}} \sum_{i=1}^{m_{tr}} (p_i - \hat{p}_i)^2$$

where, p and $\hat{p}$ are actual (or, ground-truthed) and predicted penmanship scores, m_{tr} is the number of samples in training set, θ is the set of learning parameters that our model learns after proper training.

Weight optimization. To optimize the learning weights of the model, we engaged Adam optimizer (Kingma Diederik & Adam, 2014), since it works better than some other contemporary optimizers (e.g., gradient descent, gradient descent with momentum, RMSProp; Zhang et al., 2021). In Algorithm 1, we present the Adam optimizer. The hyperparameters were fixed experimentally as follows: initial_learning_rate (α) = 10^{-3}; exponential decay rates for

the 1st and 2nd moment estimates, i.e., $\beta_1 = 0.9$, $\beta_2 = 0.999$; zero-denominator removal parameter (ε) = 10^{-8}. For training, we chose mini-batch size as 64.

Algorithm 1. Adam optimizer

Initialize: $v_{i,0} = 0;\ u_{i,0} = 0; t = 1;$

while $\theta_{i,t}$ *not converged* (*on iteration* t) ***do*** :

$$v_{i,t} = \beta_1 v_{i,t-1} + (1-\beta_1)\frac{\partial L(\theta)}{\partial \theta_{i,t}}\ ;\quad u_{i,t} = \beta_2 u_{i,t-1} + (1-\beta_2)\left(\frac{\partial L(\theta)}{\partial \theta_{i,t}}\right)^2$$

$$v_{i,t}^c = \frac{v_{i,t}}{(1-\beta_1^t)};\qquad u_{i,t}^c = \frac{u_{i,t}}{(1-\beta_2^t)}$$

$$\theta_{i,t} = \theta_{i,t-1} - \frac{\alpha}{\sqrt{u_{i,t}^c + \varepsilon}} v_{i,t}^c;$$

$$t = t + 1;$$

end while

EXPERIMENTS

To check the efficacy of our model, we performed extensive experimentations. We begin by discussing the experimental setup, followed by the experimental results.

Experimental setup

For performing the experiments, we procured a total of 39207 Chinese-character penscripts handwritten by 40 different writers, where every penscript sample was annotated by a penmanship score as mentioned earlier. Therefore, the database (DB) contains 39207 penscript samples with corresponding ground-truthed penmanship rating scores. The DB was randomly split into training (DB_{tr}) and testing (DB_t) sets with a ratio of $9:1$, where the sets were disjointed. We employed 10% data of DB_{tr} as the validation set DB_v. To reduce the overfitting of our model, we also employed a data augmentation technique (Zhang et al., 2021) on samples of DB_{tr}. For the augmentation, we randomly rotated the sample images within a range of $[0^o, 45^o]$, performed zoom-in/out by 20%, and shifted the images along with x-axis/ y-axis by 20% by filling up the shifted area using the nearest pixel.

The hyper-parameters of our model were tuned and fixed empirically over DB_v. Here, all the presented results were executed on DB_t. We performed the experiments on the TensorFlow-2 framework having Python 3.7.13 over a machine with the following configurations: Intel(R) Xeon(R) CPU @ 2.00GHz with 52 GB RAM and Tesla T4 16 GB GPU.

Results

The performance of our automated penmanship assessment model was evaluated with respect to *normalized Mean Absolute Percentage Error* (nMAPE) that is defined as follows:

$$\text{nMAPE} = \left(\frac{1}{m_t} \sum_{i=1}^{m_t} \frac{|p_i - \hat{p}_i|}{p_{max} - p_{min}} \right) \times 100\%$$

where, p and $\hat{p}$ are actual and predicted penmanship rating scores, p_{max} and p_{min} are maximum and minimum of actual penmanship ratings in the database, and m_t is the number of test set samples. In our database, $p_{max} = 9.33$ and $p_{min} = 1$. As a matter of fact, $0\% \leq \text{nMAPE} \leq 100\%$, and the lower nMAPE signifies the better performance.

As we mentioned earlier, the penmanship score $p \in \mathbb{R}$ and p is in the range of $[s_L, s_H]$, i.e., $s_L \leq p \leq s_H$. Here, $s_L = 1, s_H = 10$. We now introduce the notion of *score class* C_k with respect to the range of p. The k^{th} score class C_k contains the samples having penmanship score p in the range of k and $(k+1)$, i.e., $C_k: k \leq p \leq k+1$. The interval of the score class range may be closed/open based on the previous/next class ranges. For example, $C_L = C_1: s_L \leq p < 2$, $C_{L+1} = C_2: 2 \leq p < 3$, $C_3: 3 \leq p < 4$, …, $C_{H-1} = C_9: 9 \leq p \leq S_H$.

In Table 1, we present the performance in terms of nMAPE of the score classes, when both the actual (p) and predicted ($\hat{p}$) scores are within the corresponding classes. From this table, we can observe that our model performed the best on test samples of score class C_5 and produced 4.97% of nMAPE. For C_3, the model attained the highest nMAPE, i.e., 14.27%. Our

model achieved greater than 10% nMAPE for test samples of four classes, i.e., C_6, C_7, C_9, and C_3. Overall, we achieved 9.82% nMAPE across all the score classes.

Table 1. Model performance over various score classes

Score class	Actual (p) and predicted ($\hat{p}$) penmanship score range	nMAPE (%)
C_1	$1 \leq p, \hat{p} < 2$	6.92
C_2	$2 \leq p, \hat{p} < 3$	9.14
C_3	$3 \leq p, \hat{p} < 4$	14.27
C_4	$4 \leq p, \hat{p} < 5$	6.99
C_5	$5 \leq p, \hat{p} < 6$	4.97
C_6	$6 \leq p, \hat{p} < 7$	10.48
C_7	$7 \leq p, \hat{p} < 8$	12.67
C_8	$8 \leq p, \hat{p} < 9$	9.19
C_9	$9 \leq p, \hat{p} \leq 10$	13.72
Overall	$1 \leq p, \hat{p} \leq 10$	**9.82**

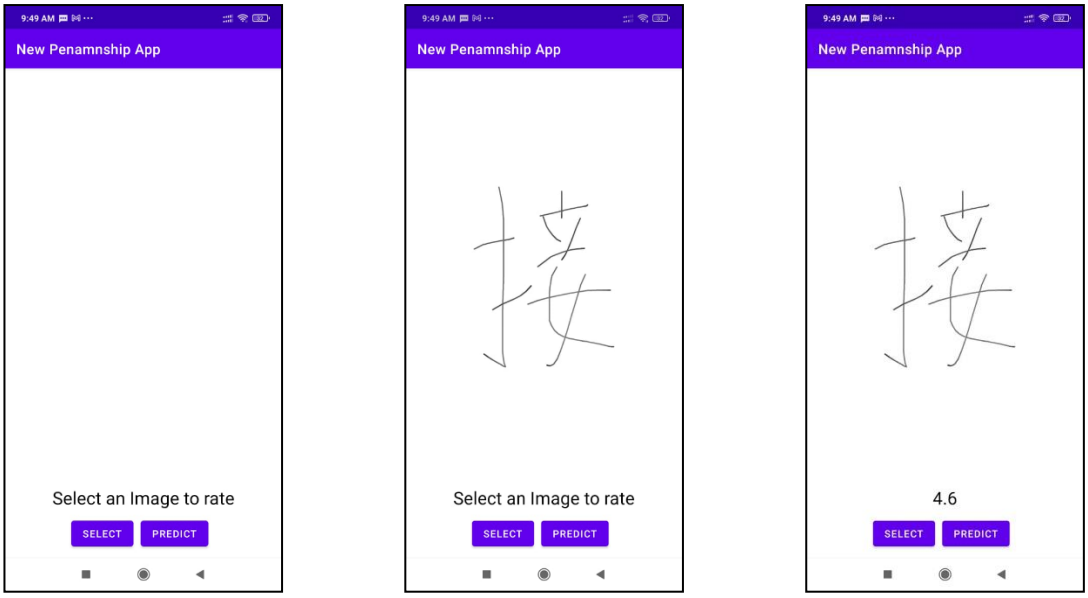
Furthermore, the score classes were categorized into low, medium, and high-scored group to have an overview on the model performance. Here, C_1, C_2, C_3 classes were joined together to form the *low*-scored group. Similarly, C_4, C_5, C_6 and C_7, C_8, C_9 formed *medium* and *high*-scored groups, respectively. The model performances on these groups are shown in Table 2. The best performance, i.e., 7.48% of nMAPE was achieved for the medium scored group, that was followed by the low and medium scored groups.

Table 2. Model performance over various score class group

Score class group	Actual (p) and predicted ($\hat{p}$) penmanship score range	nMAPE (%)
Low (C_1, C_2, C_3)	$1 \leq p_i < 4$	10.11
Medium (C_4, C_5, C_6)	$4 \leq p_i < 7$	7.48
High (C_7, C_8, C_9)	$7 \leq p_i \leq 10$	11.86

Mobile application. We also built an Android mobile application with our model embedded, so that this application can capture a penscript sample and predicts its penmanship

score in realtime (Butler, 2010). The application was built on Android Studio Bumblebee 2021.1.1 with OpenJDK 64-bit Server VM Temurin-17.0.1 installed in Windows 10 operating system. The minimum requirement of this application is SDK version 23 or Android 6.0 (Android Marshmallow). We tested the application on AVD (Android Virtual Device), i.e., Pixel 2 API 30 created within Android Studio, and two real devices, i.e., Redmi 6 Pro with Android 9 and Realme 9 with Android 12. Some screenshots of this application are shown in Figure 2.



(a) (b) (c)

Figure 2. Screen captures of the penmanship rating score prediction mobile app: (a) application screen before penscript input, (b) application screen after penscript input, (c) application screen of penmanship prediction for an input penscript.

Descriptive analysis of penmanship rating

Figure 3 shows the descriptive statistics of penmanship ratings for all the 40 writers using a boxplot (Boddy et al., 2009). One writer had an upper quartile of the penmanship score more than 7; six had an upper score quartile between 6 to 7; one had a lower score quartile below 3. There were nine writers with an upper whisker more than score 8 and nineteen writers with a lower score whisker less than 2. Overall, 32 out of the 40 writers had lower and upper quartiles within the penmanship rating score range of 3 to 6, i.e., within medium score class group. This is one possible reason for obtaining the best performance for medium-scored group (refer to Table 2).

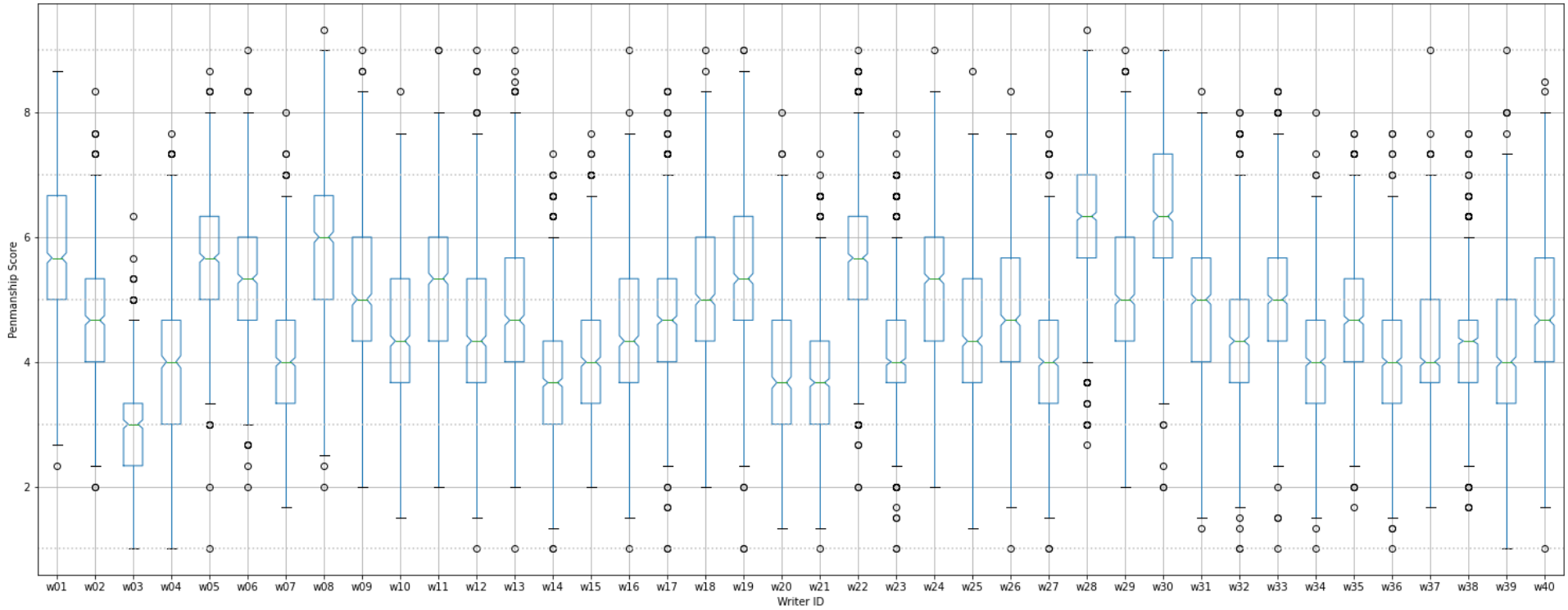


Figure 3. Boxplot of penmanship ratings for all writers' penscripts of our database. The *notch* within the box indicates the median score of a writer, the *whiskers* extending from the box represent the score variability outside the upper/lower quartiles, and the *dots* denote some possible unusual scores/outliers.

In Figure 4, we plot the number of characters/penscripts against penmanship score classes. The plot became a bell curve due to depicting the real-world scenario, which attests the integrity of our data collection. There are very few samples for class C_1 and C_9, i.e., 46 and 98, respectively. The mode penmanship score class of our dataset is C_4 that comprises 11229 samples. The score classes C_1, C_2, C_7, C_8, C_9 contain less than the mean sample count across the class. However, to perform deep regression, we required more number of samples; therefore, we employed data augmentation (Zhang et al., 2021) as mentioned earlier. It may be noted that the task could also be formulated as a classification problem; then, our model would provide a score class (i.e., a penmanship score range) instead of a specific penmanship rating score for a penscript. However, we surveyed and concluded that the general people need to know the exact penmanship score (as our application provides, refer to Figure 2); therefore, we decided to formulate the task as a regression problem. Additionally, our model can provide the score class as we presented in Table 1.

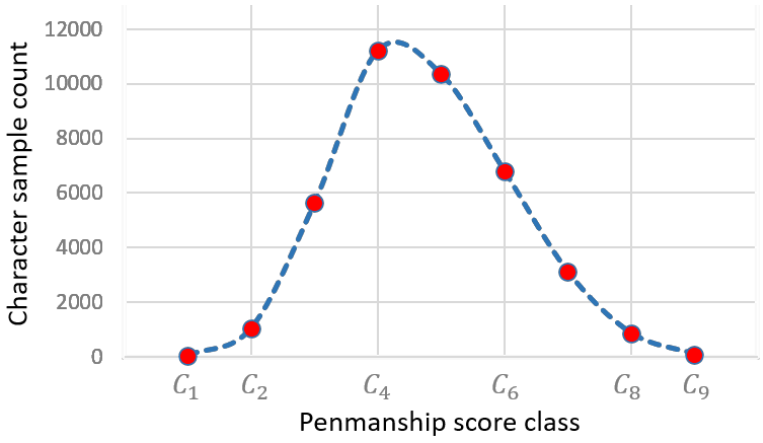


Figure 4: Sample frequency distribution over penmanship score class.

DISCUSSION

In this study, we collected human penmanship ratings for 39207 penscripts of traditional Chinese characters (comprising 1200 characters from each of 40 writers) considering both aesthetic and legibility aspects of penmanship. We then developed a CNN and trained it on these penmanship ratings. Testing on unseen penscripts revealed that the model obtains an overall 9.82% nMAPE in penmanship prediction (on a 10 point-scale). We then developed a real-time mobile application utilizing CNN to provide real-time penmanship assessment in (traditional) Chinese handwriting.

One possible explanation for the observed differences in nMAPE results across the low, medium, and high groups is the varying sample sizes. The medium group, which displays a lower nMAPE, comprises a larger number of samples (refer to Fig. 4). It is plausible that the CNN model learned more distinctive features from the larger dataset in the medium group, leading to better performance (i.e., lower nMAPE). Although we used the data augmentation techniques and have balanced the sample sizes across three groups, the synthetic nature of data augmentation, although helpful, might not fully replicate the nuances of real handwritten penscript, which could have affected the learning outcomes in the high and low groups.

We anticipate a wide range of applications of our assessment in both research and educational domains. There is a widespread concern that digital typing has significantly marginalized handwriting in written communication, leading to a decline in handwriting literacy in the Greater China area (e.g., Almog, 2019; Hilburger, 2016). Questionnaire surveys (Lan, 2013; Zhou, 2013) and empirical studies (Huang, Lin, et al., 2021; Huang, Zhou, et al., 2021) have reported that character amnesia could be attributed to fewer paper-pen practice opportunities, as a lack of handwriting practice would deteriorate explicit orthographic-motor knowledge (Christensen, 2004; Jones &

Christensen, 1999). The decline in handwriting literacy and the deterioration of penmanship in the digital age have raised concerns about the artistic expression and aesthetic beauty that can be conveyed through handwritten communication. However, research on Chinese writers' penmanship in the digital age and its potential impact on individual differences (e.g., daily exposure to handwriting) is scarce. Our assessment tool thus can be used to conveniently generate objective penmanship ratings. It is important to highlight that while our assessment is developed for penmanship evaluation of traditional character penscripts, it has the potential to be applied to simplified Chinese penscripts. Indeed, for this purpose, our model would entail fine-tuning with distinct parameters to capture the penmanship features specific to users of simplified Chinese characters. The effectiveness of this adapted model would be contingent upon the accuracy of penmanship ratings and the degree of variation in handwriting penmanship within the simplified Chinese character user population.

The CNN model, particularly the associated mobile application, can be utilised in various circumstances to assess penmanship. As individuals can use it for a quick self-assessment of their penmanship by writing one or more characters and uploading the penscript images onto the mobile application to receive penmanship ratings. To evaluate multiple penscripts, one can compute the average of the penmanship scores generated by the model/app. While our trained model primarily emphasizes the evaluation of single-character penscripts, it has the potential to assess penmanship for penscripts at the sentence or even passage level. However, the accuracy of the model in these cases still needs to be examined. The model/application can be employed to assess penmanship in various handwriting tasks, such as dictation or copying, as long as these tasks yield images of penscripts that can be used as input for the model/app.

The introduction of digital typing in schools has led to a decline in handwriting exposure for school children (Deardorff, 2011; Konnikova, 2014), which may negatively impact handwriting, particularly penmanship. Good and fluent Chinese handwriting requires years, if not decades, of handwriting practice to consolidate orthographic and visual-motor knowledge in Chinese character writing (Tong & McBride-Chang, 2010). Therefore, age could be an essential factor in the modulation of penmanship. Based on the development of current automatic penmanship assessment, future studies could collect handwriting data from different ages (e.g., primary and middle school students) to develop penmanship assessment tools that are suitable for students of different ages.

Our research also has the potential to contribute to the pre-screening of children with developmental dysgraphia, a learning disorder where children experience difficulties learning to handwrite. This disorder is estimated to affect approximately 7% to 15% of school-age children (McCloskey & Rapp, 2017). Specifically, compared to typically developing peers, children with developmental dysgraphia tend to exhibit greater variation in handwritten character size, more revised letters, and increased spatial misalignment of letters (Prunty & Barnett, 2017; Rosenblum et al., 2006; Rosenblum et al., 2004). In the context of Chinese writing, Meng et al. (2003) demonstrated that children with developmental dysgraphia often produce abnormal handwriting features, including inappropriate spacing between radicals, as well as irregular stroke lengths and connections. Consequently, children with developmental dysgraphia may receive very low penmanship ratings; a phenomenon that can be exploited (e.g., using the CNN model) for preliminary identification of children who may require clinical assessment for developmental dysgraphia. While establishing a general performance reference would necessitate gathering norms across various age

groups, the actual effectiveness of this model hinges on the differences in penmanship between typically developing individuals and those with dysgraphia across different ages. Additionally, the success of this approach also depends on our ability to precisely adjust a CNN model architecture for evaluating penmanship.

It should be noted that handwriting difficulties are multifaceted, where impairments in various cognitive processes of handwriting (such as the orthographic lexicon, semantic system, orthographic working memory, etc.) may lead to a range of abnormal handwriting performances that might not be directly related to penmanship. For example, individuals with deficiencies in the orthographic lexicon could demonstrate, in comparison to typically-developed counterparts, lower accuracy in writing less frequently-used characters (e.g., Rapp et al., 2016; Buchwald & Rapp, 2009; Rapp & Dufor, 2011). However, it is noteworthy that the penmanship in successfully executed handwriting could still appear normal.

Our model/application has the potential to identify elderly individuals who may be in the early stages of neurodegenerative diseases, as suggested by De Stefano et al. (2019) and Nackaerts et al. (2017), older adults with neurodegenerative diseases often write smaller words and have poorer writing quality compared to their healthy counterparts. Additionally, they frequently face challenges in orthographic retrieval, as noted by Rapcsak et al. (1989). These characteristics are likely to result in lower scores in a well-calibrated penmanship assessment. To effectively detect these signs, our model would require recalibration to identify specific penmanship features prevalent in the elderly population. The accuracy of diagnosing neurodegenerative diseases using this method would hinge on the extent of handwriting variation between the normal elderly and those with neurodegenerative diseases, as well as on how precisely we fine-tune the model for this purpose.

CONCLUSION

We developed and trained a CNN model for penmanship assessment in Chinese handwriting. The model was capable of providing highly human-like penmanship ratings. We further developed a mobile application that implements the CNN model to offer real-time penmanship assessment. For experiments, we created a database comprising 39207 Chinese penscript images with ground-truthed penmanship scores, where our model attained an overall 9.82% normalized Mean Absolute Percentage Error (nMAPE) on the test set, which is quite an encouraging performance. Both the CNN model and the mobile application can be employed for a wide range of academic and practical purposes.

References

Adak, C., Chaudhuri, B. B., & Blumenstein, M. (2017). Legibility and aesthetic analysis of handwriting. *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 175-182.

Adak, C., Chaudhuri, B. B., & Blumenstein, M. (2018). Cognitive Analysis for Reading and Writing of Bengali Conjuncts. *In 2018 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-7). IEEE.

Adak, C., Chaudhuri, B. B., & Blumenstein, M. (2019). An empirical study on writer identification and verification from intra-variable individual handwriting. *IEEE Access*, *7*, 24738-24758.

Adak, C., Chaudhuri, B. B., & Blumenstein, M. (2021). A Deep Reinforcement Learning-based Study on Handwriting Difficulty Analysis. *In Advances in Pattern Recognition and Artificial Intelligence* (pp. 97-117).

Alamargot, D., Morin, M.-F., & Simard-Dupuis, E. (2020). Handwriting delay in dyslexia: Children at the end of primary school still make numerous short pauses when producing letters. *Journal of learning disabilities*, *53*(3), 163-175.

Almog, G. (2019). Reassessing the evidence of Chinese "character amnesia". *The China Quarterly*, 238, 524-533.

Asselborn, T., Gargot, T., Kidziński, Ł., Johal, W., Cohen, D., Jolly, C., & Dillenbourg, P. (2018). Automated human-level diagnosis of dysgraphia using a consumer tablet. *NPJ digital medicine*, *1*(1), 1-9.

Berninger, V. W., Mizokawa, D. T., & Bragg, R. (1991). Scientific practitioner: Theory-based diagnosis and remediation of writing disabilities. *Journal of school psychology*, *29*(1), 57-79.

Boddy, R., Smith, G. (2009). *Statistical methods in practice: for scientists and technologists*. John Wiley & Sons.

Bull, R., & Stevens, J. (1979). The effects of attractiveness of writer and penmanship on essay grades. *Journal of Occupational Psychology*, *52*(1), 53-59.

Butler, M. (2010). Android: Changing the mobile landscape. *IEEE pervasive Computing*, *10*(1), 4-7.

Cai, Z. G., Xu, Z., & Yu, S. (in prep). A stroke-level database of chinese handwriting: Using OpenHandWrite with PsychoPy builder GUI for capturing handwriting processes.

Caravolas, M., Downing, C., Hadden, C. L., & Wynne, C. (2020). Handwriting legibility and its relationship to spelling ability and age: Evidence from monolingual and bilingual children. *Frontiers in Psychology*, *11*, 1097.

Chan, L., & Louie, L. (1992). Developmental trend of Chinese preschool children in drawing and writing. *Journal of Research in Childhood Education*, *6*(2), 93-99.

Chan, L., Zi Juan, C., & Lai Foon, C. (2008). Chinese preschool children's literacy development: from emergent to conventional writing. *Early Years*, *28*(2), 135-148.

Chen, X., & Kao, H. S. (2002). Visual-spatial properties and orthographic processing of Chinese characters. In *Cognitive Neuroscience Studies of the Chinese Language* (p. 175–). Hong Kong University Press, HKU.

Chow, S. M., Choy, S.-W., & Mui, S.-K. (2003). Assessing handwriting speed of children biliterate in English and Chinese. *Perceptual and Motor Skills*, *96*(2), 685-694.

Christensen, C. A. (2004). Relationship between orthographic‐motor integration and computer use for the production of creative and well‐structured written text. *British journal of educational psychology*, *74*(4), 551-564.

Cutler, L., & Graham, S. (2008). Primary grade writing instruction: A national survey. *Journal of Educational Psychology*, *100*(4), 907.

De Stefano, C., Fontanella, F., Impedovo, D., Pirlo, G., & di Freca, A. S. (2019). Handwriting analysis to support neurodegenerative diseases diagnosis: A review. *Pattern recognition letters*, *121*, 37-45.

Deardorff, J. (2011). The many health perks of good handwriting. *Chicago Tribune (1963)*.

Eidlitz-Neufeld, M. R. e. (2003). *Early letter form errors as a predictor of later literacy outcomes and the short-and long-term benefits of early instruction in proper letter formation* (pp. 4218-4218). Doctoral dissertation, University of Toronto]. Toronto. http://ovidsp. ovid. com/ovidweb. cgi.

Erez, N., & Parush, S. (1999). The Hebrew handwriting evaluation School of Occupational Therapy. *Faculty of Medicine. Hebrew University of Jerusalem, Israel*.

Fairbank, A. (2018). Handwriting manual. Courier Dover Publications.

Falk, T. H., Tam, C., Schellnus, H., & Chau, T. (2011). On the development of a computer-based handwriting assessment tool to objectively quantify handwriting proficiency in children. *Computer methods and programs in biomedicine*, *104*(3), e102-e111.

Feder, K. P., & Majnemer, A. (2007). Handwriting development, competency, and intervention. *Developmental Medicine & Child Neurology*, *49*(4), 312-317.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Hilburger, C. (2016). Character amnesia: an overview. *In Sino-Platonic Papers* (Vol. 264, pp. 51-70).

Huang, S., Lin, W., Xu, M., Wang, R., & Cai, Z. (2021). EXPRESS: On the tip of the pen: Effects of character-level lexical variables and handwriter-level individual differences on orthographic retrieval difficulties in Chinese handwriting. *Quarterly Journal of Experimental Psychology*, 17470218211004385.

Huang, S., Zhou, Y., Du, M., Wang, R., & Cai, Z. G. (2021). Character amnesia in Chinese handwriting: A mega-study analysis. *Language Sciences*, *85*, 101383.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International conference on machine learning*, (pp. 448-456).

James, K. H., & Engelhardt, L. (2012). The effects of handwriting experience on functional brain development in pre-literate children. *Trends in neuroscience and education*, *1*(1), 32-42.

Jones, C., & Hall, T. (2013). The importance of handwriting: Why it was added to the Utah Core Standards for English language arts. *The Utah Journal of Literacy*, *16*(2), 28-36.

Jones, D., & Christensen, C. A. (1999). Relationship between automaticity in handwriting and students' ability to generate written text. *Journal of Educational Psychology*, *91*(1), 44.

Kingma Diederik, P., & Adam, J. B. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Konnikova, M. (2014). What's lost as handwriting fades. *The New York Times*, *2*.

Lam, S. S., Au, R. K., Leung, H. W., & Li-Tsang, C. W. (2011). Chinese handwriting performance of primary school children with dyslexia. *Research in developmental disabilities*, *32*(5), 1745-1756.

Lan, S. (2013). "Picking up a pen and forgetting characters" What have we forgotten? Difficulties in reading Chinese characters (“提笔忘字”我们究竟忘掉了什么？看汉字之困). from. http://edu.people.com.cn/n/2013/1104/c1053-23420614.html. (Accessed 6 March 2020).

Lathuilière, S., Mesejo, P., Alameda-Pineda, X., & Horaud, R. (2019). A comprehensive analysis of deep regression. *IEEE transactions on pattern analysis and machine intelligence*, *42*(9), 2065-2081.

Li, J., Liu, Y., Wang, Y., Wang, N., Ji, Y., Wei, T., ... & Yang, Y. (2023). Functional brain networks underlying the interaction between central and peripheral processes involved in Chinese handwriting in children and adults. *Human Brain Mapping*, 44(1), 142-155.

Li-Tsang, C. W., Li, T. M., Yang, C., Leung, H. W., & Zhang, E. W. (2022). Evaluating Chinese Handwriting Performance of Primary School Students Using the Smart Handwriting Analysis and Recognition Platform (SHARP). *medRxiv*.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

Longcamp, M., Boucard, C., Gilhodes, J.-C., & Velay, J.-L. (2006). Remembering the orientation of newly learned characters depends on the associated writing knowledge: A comparison between handwriting and typing. *Human movement science*, *25*(4-5), 646-656.

Longcamp, M., Tanskanen, T., & Hari, R. (2006). The imprint of action: Motor cortex involvement in visual perception of handwritten letters. *NeuroImage*, *33*(2), 681-688.

Martínez-García, C., Afonso, O., Cuetos, F., & Suárez-Coalla, P. (2021). Handwriting production in Spanish children with dyslexia: spelling or motor difficulties? *Reading and writing*, *34*(3), 565-593.

McCloskey, M., & Rapp, B. (2017). Developmental dysgraphia: An overview and framework for research. *Cognitive neuropsychology*, *34*(3-4), 65-82.

Medwell, J., & Wray, D. (2008). Handwriting–A forgotten language skill? *Language and education*, *22*(1), 34-47.

Meng, X., Zhou, X., & Wu, J. (2003). Developmental coordination disorder and dysgraphia: A case study. *Acta Psychologica Sinica*.

Mueller, P. A., & Oppenheimer, D. M. (2014). The pen is mightier than the keyboard: Advantages of longhand over laptop note taking. *Psychological Science*, *25*(6), 1159-1168.

Nackaerts, E., Heremans, E., Smits-Engelsman, B. C., Broeder, S., Vandenberghe, W., Bergmans, B., & Nieuwboer, A. (2017). Validity and reliability of a new tool to evaluate handwriting difficulties in Parkinson's disease. *PloS one*, *12*(3), e0173157.

Olsen, J., & Knapton, E. (2006). The Print Tool: The tool to evaluate and remediate. *Cabin John, MD: Handwriting Without Tears*.

Peverly, S. T., Vekaria, P. C., Reddington, L. A., Sumowski, J. F., Johnson, K. R., & Ramsay, C. M. (2013). The relationship of handwriting speed, working memory, language comprehension and outlines to lecture note‐taking and test‐taking among college students. *Applied Cognitive Psychology*, *27*(1), 115-126.

Phelps, J., Stempel, L., & Speck, G. (1985). The children's handwriting scale: A new diagnostic tool. *The Journal of Educational Research*, *79*(1), 46-50.

Prunty, M., & Barnett, A. L. (2017). Understanding handwriting difficulties: A comparison of children with and without motor impairment. *Cognitive neuropsychology*, *34*(3-4), 205-218.

Rapcsak, S. Z., Arthur, S. A., Bliklen, D. A., & Rubens, A. B. (1989). Lexical agraphia in Alzheimer's disease. *Archives of Neurology*, *46*(1), 65-68.

Reisman, J. (1999). *Minnesota handwriting assessment*. Psychological Corporation.

Reisman, J. E. (1993). Development and reliability of the research version of the Minnesota Handwriting Test. *Physical & Occupational Therapy in Pediatrics*, *13*(2), 41-55.

Rosenblum, S., Dvorkin, A. Y., & Weiss, P. L. (2006). Automatic segmentation as a tool for examining the handwriting process of children with dysgraphic and proficient handwriting. *Human movement science*, *25*(4-5), 608-621.

Rosenblum, S., Weiss, P. L., & Parush, S. (2004). Handwriting evaluation for developmental dysgraphia: Process versus product. *Reading and writing*, *17*(5), 433-458.

Sang, Y. (2023). Uncovering language socialization mechanisms in language teacher identity formation: An ethnographic study in a Chinese culture class. *Linguistics and Education*, 73, 101138.

Sheffield, B. (1996). Handwriting: A neglected cornerstone of literacy. *Annals of Dyslexia*, *46*(1), 21-35.

Simner, M. L. (1988). Predicting First Grade Achievement from Form Errors in Printing at the Start of Pre-Kindergarten. *ERIC*

Skar, G. B., Lei, P.-W., Graham, S., Aasen, A. J., Johansen, M. B., & Kvistad, A. H. (2022). Handwriting fluency and the quality of primary grade students' writing. *Reading and writing*, *35*(2), 509-538.

Tong, X., & McBride-Chang, C. (2010). Developmental models of learning to read Chinese words. *Developmental psychology*, *46*(6), 1662.

Tse, L. F. L., Siu, A. M. H., & Li-Tsang, C. W. P. (2019). Assessment of early handwriting skill in kindergarten children using a Chinese name writing test. *Reading and writing*, *32*(2), 265-284.

Tseng, M.-h. (1993). Factorial validity of the Tseng handwriting problem checklist. *Journal of the Occupational Therapy Association of the Republic of China*, *11*, 13-26.

Tseng, M. H., & Hsueh, I. P. (1997). Performance of school‐aged children on a Chinese handwriting speed test. *Occupational Therapy International*, *4*(4), 294-303.

Wollscheid, S., Sjaastad, J., & Tømte, C. (2016). The impact of digital devices vs. Pen (cil) and paper on primary school students' writing skills–A research review. *Computers & education*, *95*, 19-35.

Wood, D. M. (1982). Bringing Chinese Culture Alive through Language. *The Social Studies*, 73(4), 155-159.

Yang, Y., Zuo, Z., Tam, F., Graham, S. J., Li, J., Ji, Y., . . . Ou, J. (2022). The brain basis of handwriting deficits in Chinese children with developmental dyslexia. Developmental science, 25(2), e13161.

Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2021). Dive into deep learning. *arXiv preprint arXiv:2106.11342*.

Zhou, Y., Xu, J.,. (2013). 98.8% of respondents have forgotten to write (98.8%受访者曾提笔忘字). from. http://zqb.cyol.com/html/2013-08/27/nw.D110000zgqnb_20130827_2-07.htm. (Accessed 6 March 2020).

Kiefer, M., & Velay, J. L. (2016). Writing in the digital age. *Trends in Neuroscience and Education*, 5(3), 77-81.

Marquardt, C., Meyer, M. D., Schneider, M., & Hilgemann, R. (2016). Learning handwriting at school–A teachers' survey on actual problems and future options. *Trends in Neuroscience and Education*, 5(3), 82-89.