A ResNet-Based Deep Learning Model for Automated Scoring of Elementary Students' Chinese Calligraphy

Yi Pei Lin, National Taiwan Normal University, Taiwan
Tzren-Ru Chou, National Taiwan Normal University, Taiwan

The Southeast Asian Conference on Education 2026
Official Conference Proceedings

Abstract

This study proposes an automatic scoring approach for calligraphy images using the deep learning model ResNet. The evaluation criteria are grounded in calligraphic aesthetics, emphasizing stroke techniques and the holistic relationships of character structure. The dataset consists of elementary students' handwriting samples and corresponding expert ratings. After preprocessing and model training, the system generates predicted scores. Experimental results show that the model achieves a low mean absolute error (MAE = 2.6) and a high quadratic weighted kappa (QWK = 0.898), indicating strong consistency between the automated scoring and expert evaluations. The proposed system provides real-time feedback and reduces teachers' assessment workload. Future work will expand the dataset and refine the scoring dimensions to enhance the model's applicability in calligraphy education.

Keywords: ResNet model, convolution, automatic calligraphy evaluation, elementary visual arts education

iafor
The International Academic Forum
www.iafor.org

Introduction

Calligraphy education plays an important role in elementary visual arts and language learning by cultivating cultural understanding and aesthetic perception through brushwork and structural balance. According to the Ministry of Education (2018a, 2018b), calligraphy is integrated into both Language Arts and the Arts domain within the 12-year basic education curriculum guidelines to enhance students' aesthetic literacy. However, as noted in the Ministry of Education's (2024) latest medium-term plan for aesthetic education, promoting calligraphy in modern classrooms faces challenges due to the subjectivity of assessment. Yang (2009a) identifies various dilemmas in elementary calligraphy teaching, particularly the inconsistency in feedback, while Yang (2009b) emphasizes the need for objective assessment tools to measure students' progress accurately.

Historically, Chinese calligraphy has evolved through distinct periods, with works from the late Qing to the Japanese colonial period reflecting significant social changes in Taiwan (Huang, 2010). Foundational theories, such as those discussed in Sun's (2003) history of Chinese calligraphy and Kang's (1920) classic treatise, provide the aesthetic standards used for evaluation today. In recent years, researchers have attempted to bridge art and science. Fu (2010a, 2010b) provides principles for the scientific analysis of calligraphy, while Fu (2010c) explores the artistic thoughts behind Tang Dynasty masterpieces.

Research Methodology

The evaluation criteria for this study are grounded in established calligraphic theories. We specifically focus on the structural configurations characterized by masters like Ouyang Xun (Peking University Calligraphy Research Center, 2015a, 2015b). Studies on the "Jiucheng Palace Sweet Spring Inscription" by Xu (2010a) and the structural analysis of Ouyang Xun's style by Xu (2010b) inform our model's feature extraction layer. Furthermore, the design of our assessment rubrics incorporates the curriculum auxiliary manuals provided by Qiu et al. (2006) and the grading standards for junior high school calligraphy developed by Cai (2015).

The dataset acquisition was influenced by current regional educational practices, such as the student calligraphy competitions organized by the Cultural Affairs Bureau of Taichung City Government (2023). This ensures that the training data reflects contemporary student handwriting levels. The methodology also considers the historical influence of calligraphy exhibitions and societies, as explored by Xu (2011), to ensure the model accounts for stylistic variations.

Calligraphy Samples and Participants

This stage involved the systematic collection of 2,000 physical calligraphy samples, all executed on standardized paper with a uniform grid structure to ensure experimental consistency. These physical works were subsequently digitized into high-resolution PDF format, establishing a robust empirical foundation with the structural integrity required for subsequent computer vision analysis.

CNN-Based Automated Evaluation Module Design

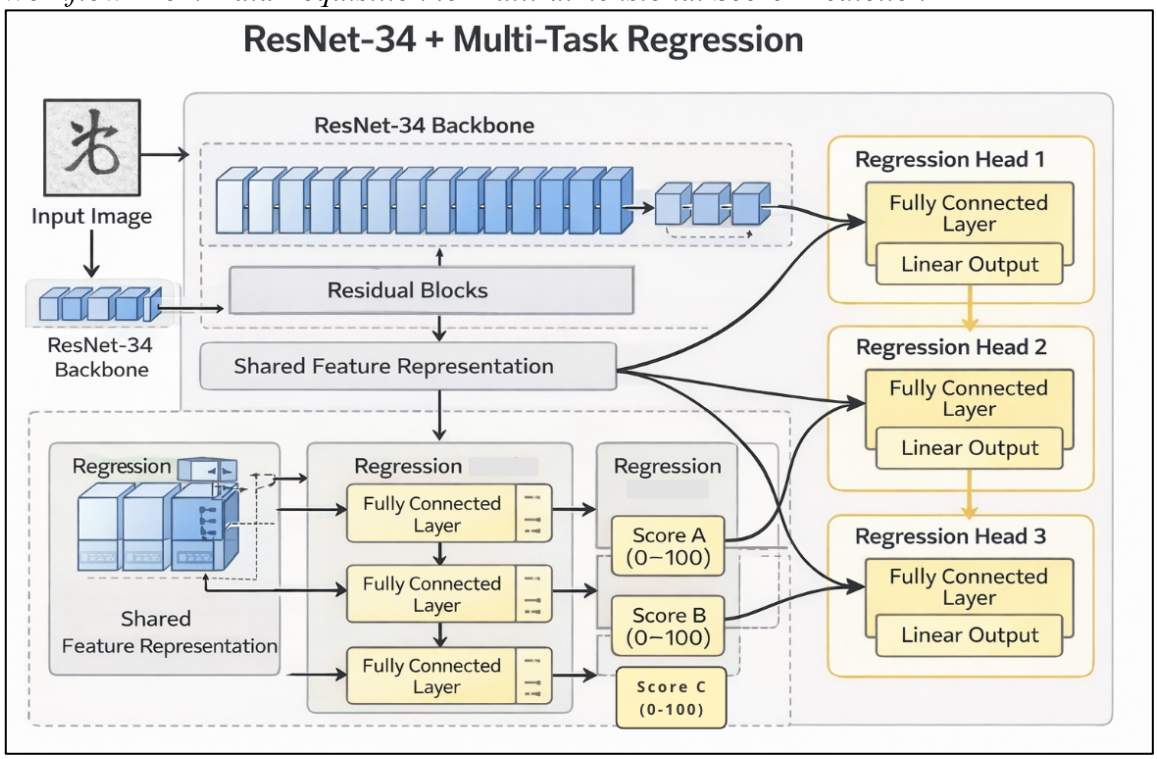
The proposed methodology is structured into three integrated phases: image preprocessing, feature learning, and model execution. Initially, scanned calligraphy images were converted to

grayscale and normalized to eliminate color-induced variance, with each character resized to a fixed resolution and centrally aligned to ensure visual consistency across the dataset. Subsequently, a convolutional neural network (ResNet-34) was employed to extract hierarchical visual features, enabling the model to learn representations ranging from low-level stroke patterns to high-level structural configurations.

This ResNet-34 backbone was fine-tuned on the calligraphy dataset to perform regression, outputting continuous score predictions based on the identified features. To enhance model stability and reduce training noise, visually ambiguous or low-quality calligraphy samples were excluded from the final training set, ensuring the reliability of the automated evaluation framework.

Figure 1

The Proposed Automated Calligraphy Evaluation Framework, Illustrating the Four-Stage Workflow From Data Acquisition to Multi-dimensional Score Prediction



Expert Manual Evaluation

The establishment of the ground truth involved three calligraphy experts, each possessing extensive pedagogical experience. The evaluation process was conducted through an independent scoring protocol, where each expert appraised the same dataset of calligraphy samples based on a set of predefined standardized criteria. To ensure objectivity and minimize individual bias, the scores were subsequently aggregated, and their arithmetic mean was adopted as the authoritative reference standard (ground truth) for both model training and performance benchmarking.

Consistency Assessment: QWK and MAE

Model efficacy was determined by comparing predicted scores against expert ratings. Mean Absolute Error (MAE) quantified prediction deviation, while Quadratic Weighted Kappa

(QWK) examined the statistical consistency and categorical alignment between the AI outputs and expert evaluations.

This dual-metric approach ensures both the accuracy and pedagogical validity of the system.

Results

Based on the 2,000 calligraphy samples evaluated in this study, the average Quadratic Weighted Kappa (QWK) between the model-predicted scores and expert ratings was approximately 0.75. In some cases, the QWK exceeded 0.80, indicating that the proposed ResNet-34–based model can produce evaluations closely aligned with expert judgment in elementary calligraphy assessment.

A small number of samples showed larger discrepancies between model predictions and expert scores, mainly due to subtle variations in stroke execution or complex structural arrangements within characters. In a few cases, irregular writing patterns or visually ambiguous features made consistent evaluation more difficult. Although automated evaluation can substantially reduce teachers' assessment workload, these results underscore the importance of expert review and refinement to ensure reliable assessment quality.

Table 1
Development of a Multi-Dimensional Calligraphy Evaluation System Using ResNet-34 Deep Learning Architecture

Dataset	N	QWK	MAE
Calligraphy samples	2000	0.75	2.63

Conclusion

This study proposed a ResNet-34–based automated calligraphy evaluation method for intermediate elementary students, using deep learning to predict scores from calligraphy images under an educational assessment framework. Many model predictions showed close agreement with expert ratings, effectively reducing teachers' evaluation workload while maintaining assessment consistency.

Although the model may still exhibit limitations when handling visually ambiguous writing patterns or highly complex structural arrangements, further refinement of training data and feature learning strategies is expected to improve performance. Rather than replacing expert judgment, the proposed approach functions as an assistive assessment tool, enhancing efficiency and reliability in classroom practice. Future work may extend this framework to larger datasets, additional script styles, and multi-level feature analysis, further supporting the integration of artificial intelligence into art education assessment.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

The author declares that Gemini, an AI-assisted writing software, was used in proofreading and refining the language used in the manuscript. The usage was limited to correcting grammatical and spelling errors and rephrasing statements for accuracy and clarity. The author further

declares that, apart from Gemini, no other AI or AI-assisted technologies have been used to generate content in writing the manuscript. The ideas, design, procedures, findings, analyses, and discussion are originally written and derived from careful and systematic conduct of the research.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® v.2. *Journal of Technology, Learning, and Assessment, 4*(3), 1–30.

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. ASCD.

Cai, X. (2015). *The construction of assessment rubrics for visual arts in junior high school: A case study of calligraphy art* [Unpublished master's thesis]. National Taiwan University of Arts.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213–220.

Cultural Affairs Bureau of Taichung City Government. (2023). *The 17th Dadun Cup student calligraphy competition* [Press release].

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805.

Fu, S. (2010a). *Principles of scientific analysis of calligraphy art (Vol. 1)*. Art & Collection Group.

Fu, S. (2010b). *Principles of scientific analysis of calligraphy art (Vol. 2)*. Art & Collection Group.

Fu, S. (2010c). Sun Guoting's "Shupu" and artistic thought of the Tang Dynasty. *Archives of Asian Art, 60*(1), 25–49.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).

Huang, H. (2010). An investigation of Taiwan's calligraphy under social changes from the late Qing to the Japanese colonial period (1885–1945). In *Proceedings of the symposium on the retrospective of Taiwan calligraphy development in the 20th century* (pp. 183–212). Taipei Physical Education College.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679–688.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*, 1097–1105.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2–3), 259–284.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.

Ministry of Education. (2018a). *Curriculum guidelines of 12-year basic education: Language arts.*

Ministry of Education. (2018b). *Curriculum guidelines of 12-year basic education: Arts.*

Ministry of Education. (2024). *The third phase of the medium and long-term plan for aesthetic education (2024–2028).*

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 47*(5), 238–243.

Peking University Calligraphy Research Center. (2015a). *Ouyang Xun calligraphy theory database*. Peking University.

Peking University Calligraphy Research Center. (2015b). *On the history of Chinese calligraphy*. Peking University.

Qiu, Y., Zhang, S., & Li, X. (2006). *Curriculum auxiliary teaching reference manual 3 for senior high school art: Basic design*. National Taiwan Arts Education Center.

Shermis, M. D. (2018). *Contrasting state-of-the-art automated scoring of essays*. Routledge.

Simonyan, K., & Zisserman, A. (2015). *Very deep convolutional networks for large-scale image recognition*. arXiv preprint arXiv:1409.1556.

Sun, J. (2003). *The history of Chinese calligraphy*. Wu-Nan Book Inc.

Talebi, H., & Milanfar, P. (2018). NIMA: Neural image assessment. *IEEE Transactions on Image Processing, 27*(8), 3998–4011.

Vaswani, A., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing, 13*(4), 600–612.

Xu, F. (2011). *The calligraphy exhibition of the ten-person calligraphy society and its influence* [Unpublished master's thesis]. Huafan University.

Xu, Z., Mittal, P. S., Ahmed, M. M., Adak, C., & Cai, Z. G. (2025). Assessing penmanship of Chinese handwriting: A deep learning-based approach. *Reading and Writing, 38*, 723–743.

Yang, F. (2009a). *A study on the dilemmas and strategies of calligraphy teaching in elementary schools* [Unpublished master's thesis]. National Hsinchu University of Education.

Zhou, Y., Wang, L., & Li, X. (2022). Deep learning for Chinese calligraphy evaluation and classification. *Journal of Cultural Heritage, 54*, 213–225.

Contact email: puddingding11@gmail.com