Scientific and Social Research, 2025, Volume 7, Issue 8
https://ojs.bbwpublisher.com/index.php/SSR
Online ISSN: 2981-9946
Print ISSN: 2661-4332

Research and Practice of Multi-dimensional Chinese Character Quantification Evaluation Methods

Peng Deng*, Guiying Yang

Chongqing Energy College, Chongqing 402260, China

**Author to whom correspondence should be addressed.*

Abstract: Addressing the pain points of traditional Chinese character evaluation, which is highly subjective and lacks quantitative standards, this paper proposes a three-dimensional intelligent quantitative evaluation model for Chinese characters based on "accuracy—balance—standardization", transitioning from experience-driven to data-driven evaluation: improving multi-model convolutional neural networks to extract stroke keypoints for accuracy assessment, combining image moments with cosine similarity to measure center of gravity deviation for balance evaluation, and using 3×3 grid pixel density ratios to characterize structural framework for standardization measurement. The model operates in a closed-loop system of "recognition—quantification—fuzzy comprehensive evaluation", enhancing recognition accuracy through hard voting, automatically calibrating the scoring range using the "best-worst sample set", and obtaining intelligent weights of 39:32:29 through backpropagation training. Experiments on four types of samples from the CASIA-HWDB1.1 and custom datasets show that the average error between the model and expert scores under intelligent weights is significantly lower than that of the two fixed weights, providing a practical quantitative evaluation tool for Chinese character education and cultural heritage.

Keywords: Multi-dimensional evaluation framework; Convolutional neural network; Fuzzy comprehensive evaluation; Writing quality assessment

Online publication: September 17, 2025

1. Introduction

As the only ancient writing system in the world that has continued to the present day, Chinese characters are also the language with the largest number of users today. With the ongoing globalization of communication, the popularity and demand for learning Chinese characters are growing steadily. As a carrier of China's millennia-old civilization, Chinese characters are characterized by their diverse categories, complex structural components, and numerous homophones. For beginners, it is not only necessary to master the stroke order and

morphological characteristics of Chinese characters, but also to understand their internal structure. For example, subtle differences between characters that look similar can lead to significant differences in their meaning and usage, thereby increasing the difficulty of learning Chinese characters. At the same time, traditional Chinese character evaluation often focuses solely on writing standards or aesthetic dimensions, which can be influenced by personal experience and preferences. This leads to significant scoring discrepancies for the same character, making it difficult to establish an objective evaluation system. Such inconsistencies can have a negative impact on learners' motivation. Additionally, timely evaluation during the process of learning Chinese characters is another major obstacle. Learners who cannot receive evaluations of their Chinese characters within a limited timeframe may see their interest and efficiency in learning affected. The primary cause of these issues lies in the lack of a unified, quantifiable, systematic, and comprehensive evaluation system. Therefore, the objective of this study is to construct a multi-feature integrated Chinese character intelligent evaluation model. By integrating neural network technology with traditional techniques, it combines structural evaluation and aesthetic evaluation to form a multi-dimensional quantitative evaluation system, providing a referenceable quantitative evaluation approach for the research and learning of Chinese characters.

2. Current status of Chinese character evaluation research

With the widespread use of smart devices, the phenomena of "forgetting how to write characters" and poor writing quality have become commonplace. Additionally, learners' laziness due to the lack of timely evaluation tools is a common contributing factor to these issues. These problems have a certain negative impact on the inheritance of Chinese culture, prompting researchers to dedicate themselves to studying the recognition and evaluation of Chinese characters to contribute to cultural preservation. Currently, research on Chinese character evaluation primarily focuses on aspects such as structure, writing standardization, and stroke quality, but lacks a unified quantitative standard [1]. Based on previous research, the research directions can be categorized into three types: rule-based evaluation, similarity-based evaluation, and machine learning-based evaluation.

For rule-based evaluation methods, Zhuang Chongbiao et al. proposed a stroke-based grading strategy, establishing normative rules for stroke-to-stroke features. However, this method relies on manually defined error types and cannot judge undefined error types, lacking comprehensiveness [2]. Han Ruifang et al. conducted a hierarchical classification of stroke keypoints and stroke segment details, enabling the evaluation of structural features. However, this method requires a high level of expertise in evaluating stroke features [3]. In summary, these stroke-rule-based methods have laid the foundation for subsequent research and achieved preliminary success in Chinese character evaluation studies. However, they heavily rely on rules and lack good universality.

Building on the shortcomings of rule-based evaluation methods, researchers have proposed using similarity-based calculation methods to enhance universality. Such methods typically involve first extracting features from the sample to be tested, then comparing them with a template to calculate similarity, and finally deriving an evaluation of writing quality based on the similarity score. Yu Kai et al., as one of the pioneers in similarity feature research, proposed calculating similarity using skeleton and contour features, enabling rapid matching and evaluation even with large datasets [4]. Wu Chuzhou et al. further optimized the skeleton feature evaluation approach by proposing a method that involves grid partitioning of the skeleton followed by matrix calculations [5]. Ge Jiamin proposed using fuzzy probability distribution calculations to define a range for evaluating writing quality [6]. She used fuzzy comprehensive evaluation to calculate the expected value range for individual characters, then compared the

expected values of the test samples with the expected value range to derive the evaluation results. However, while these studies fully utilize Chinese character features and achieve good results, their generalization capabilities still need improvement. Some studies have attempted to introduce machine learning methods for automatic evaluation of writing quality. For example, Zhuang Ziming et al. proposed using convolutional neural networks to extract features and retrieve images with high similarity scores based on similarity matching, thereby achieving the evaluation objective [7]. Although this approach demonstrated good evaluation results, it relies heavily on image annotation. This paper combines machine learning features with the structural features of Chinese characters to explore the effects of fuzzy comprehensive evaluation.

3. Design of a multi-feature fusion Chinese character quantification evaluation model

3.1. Determination of evaluation dimensions

Unlike numbers and letters, Chinese characters are characterized by their diverse categories and complex structures. Traditional evaluations of Chinese characters primarily rely on manual assessment, which is essentially a post-hoc evaluation. When evaluating the aesthetic quality of Chinese characters, teachers primarily focus on elements such as stroke execution, position, and character shape, which inherently involve a significant degree of subjectivity. Given the general applicability of evaluation strategies and the issues mentioned above, this paper employs a Convolutional Neural Network (CNN) to address accuracy issues and extracts CNN features as one of the evaluation dimensions to measure the accuracy of writing. This is because multi-layer convolutional operations can extract deep visual features such as stroke details of Chinese characters, eliminating the need for manually designed features while simulating human visual perception of the aesthetic quality of Chinese characters. Secondly, the center of gravity and grid features of Chinese characters are selected. The center of gravity of each Chinese character is a fundamental element in evaluating aesthetic appeal. It can be used to quantitatively assess whether a Chinese character has "imbalance" issues (such as tilting or misalignment), as the center of gravity distribution of a properly written Chinese character is stable and harmonious. Therefore, the basic aesthetic appeal of a Chinese character can be measured based on the difference in its center of gravity. Grid features can be used to divide Chinese characters into regions based on their center of gravity features, thereby quantitatively assessing the uniformity of stroke distribution and aligning with the "structural framework" of Chinese characters. This allows for calculations of component proportions and "white space" issues, combining macro-level structure with micro-level details. Ultimately, CNN features capture details, center of gravity features measure macro-level structure, and grid features serve as the link between macro- and micro-levels. Through fuzzy comprehensive evaluation, a comprehensive aesthetic evaluation is formed.

3.2. Design of quantitative evaluation indicators

Based on the determination of evaluation dimensions, this paper proposes to use three types of quantitative indicators for evaluation, mainly including the "correctness" indicator based on CNN features. This indicator extracts detailed features such as stroke key points and ultimately obtains the coverage of detailed features. When the proportion of detailed features of a test sample exceeds the threshold, the corresponding category can be determined. The calculated feature value is used as the micro-level aesthetic score of the sample, denoted as $Grade_{CNN}$.

The "balance" metric, based on center of gravity features, quantifies the overall balance of a Chinese

character. This is achieved by calculating the center of gravity coordinates of the sample being tested, then determining the degree of deviation from the expected sample's center of gravity, thereby quantifying the aesthetic quality of the overall structure. After preprocessing the image, it undergoes grayscale conversion and binarization. The center of gravity coordinates are then calculated using the moment of the binary image. The specific principle of metric quantification is as follows: Let the pixel coordinates of the target region in the image be (x, y), and the pixel value be I(x, y) (I(x, y) = 1 indicates the pixel exists, and 0 indicates a white background). Formula 1 is used to calculate the 0th-order moment m_{00}, and formulas 2 and 3 are used to calculate the 1st-order moments m_{01} and m_{10}, respectively. The center of gravity coordinates $(\bar{x},\bar{y})$ of the sample are calculated using formulas 4 and 5. Finally, the cosine similarity between the test sample and the target sample (where the mean centroid coordinates of the optimal sample set in the target sample are assumed to be (x_1, y_1), and the mean center of gravity coordinates of the worst sample set are (x_2, y_2). The cosine similarity is normalized in formula 6 to transform the value range to 0-1 before being used for subsequent calculations, forming a "balance" indicator score, denoted as $Grade_{gravity}$.

$$m_{00}=\sum_x \sum_y I(x,y) \quad \text{Formula 1}$$

$$m_{01}=\sum_x \sum_y x\cdot I(x,y) \quad \text{Formula 2}$$

$$m_{10}=\sum_x \sum_y y\cdot I(x,y) \quad \text{Formula 3}$$

$$\bar{x}=\frac{m_{10}}{m_{00}} \quad \text{Formula 4}$$

$$\bar{y}=\frac{m_{01}}{m_{00}} \quad \text{Formula 5}$$

$$Grade_{\text{gravity}} = \left(1-\left(1-\frac{(\bar{x},\bar{y})\cdot(x_1,y_1)}{\|(\bar{x},\bar{y})\|\cdot\|(x_1,y_1)\|}\right)\Big/\left(1-\frac{(x_1,y_1)\cdot(x_2,y_2)}{\|(x_1,y_1)\|\cdot\|(x_2,y_2)\|}\right)\right)\times 100 \quad \text{Formula 6}$$

The "normativity" metric based on grid features primarily quantifies the uniformity of stroke distribution and component proportions. Each sample is divided into a 3×3 grid, and the pixel distribution of each grid is calculated. This is then compared with the expected sample to determine whether the stroke layout and structure of the sample are normative. The specific quantification principle of the metrics is as follows: Let the sample height be H, width be W, and grid size be grid_size × grid_size (set to 3×3 in this study), each grid has a size of region_h × region_w, a pixel density of region_density, a background pixel count of region_foreground (pixels with a value of 0), and a total pixel count of region_total. The size data and pixel density of each grid are calculated using formulas 7–9. Subsequently, the pixel density matrix of all grids in the sample under test is flattened and denoted as $Grid_t$. The pixel density average matrix of the optimal sample set of the reference sample is flattened and denoted as $Grid_s$, and the pixel density average matrix of the worst sample set is flattened and denoted as $Grid_b$. Finally, the final score is calculated by substituting $Grid_t$, $Grid_b$, and $Grid_s$ into Formula 6, and it is used as the "normativity" indicator score, denoted as $Grade_{grid}$.

$$\text{region\_h} = \lfloor H/\text{grid\_size} \rfloor \quad \text{Formula 7}$$

$$\text{region\_w} = \lfloor W/\text{grid\_size} \rfloor \quad \text{Formula 8}$$

$$\text{region\_density} = \text{region\_foreground}/\text{region\_total} \quad \text{Formula 9}$$

3.3. Quantitative evaluation model design

In the process of quantitative evaluation, the prerequisite for achieving accurate and objective evaluation results is correct writing. Therefore, a multi-feature fusion recognition module is incorporated into the quantitative evaluation model to complete the recognition task. After accurate recognition, quantitative feature scoring calculations are performed. The overall model diagram is shown in **Figure 1**.

Figure 1. Quantitative evaluation model

The model recognition task is used to evaluate the correctness of Chinese characters. The model used is based on the VGG16 model, with the addition of a BatchNormalization layer to improve model training speed and convergence stability. It also incorporates the Inception feature from GoogLeNet, where the Inception submodule is adjusted based on the principles of the basic LeNet network. This ultimately forms the network model structure, a multi-model described in this paper, as shown in **Figure 2**. A hard voting mechanism (simulating multiple experts voting, with the majority vote prevailing) is integrated into the recognition model to further improve classification accuracy.

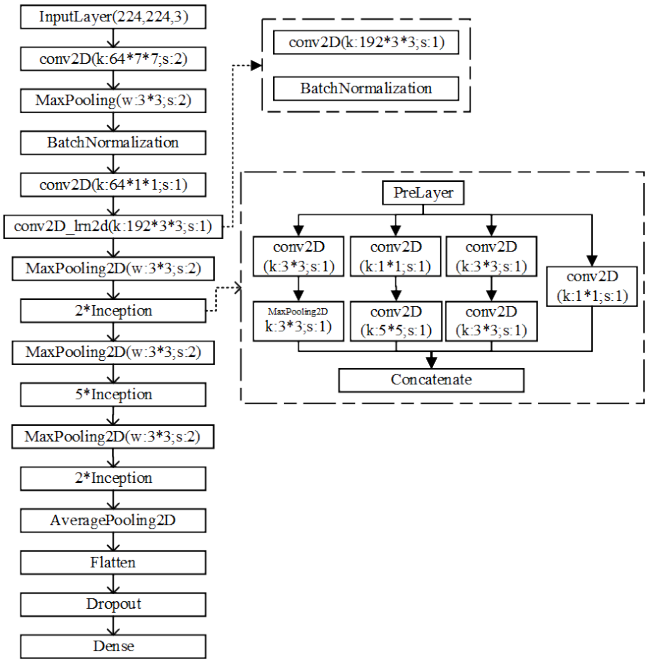

Figure 2. Multi-model

The core of the quantitative evaluation model is the calculation of quantitative features, which mainly includes three submodules: CNN feature calculation for the "correctness" indicator, center of gravity feature calculation for the "balance" indicator, and grid feature calculation for the "standardization" indicator. The CNN feature score calculation process is as follows:

1. Input the test sample into the recognition model to obtain the classification results;
2. Based on the classification results, retrieve the data and extract the 10 optimal samples with the highest probability values in the dataset to form the optimal sample set D_{good}, and the 10 samples with the lowest probability values to form the worst sample set D_{bad};
3. Calculate the average probability values of the optimal sample set and the worst sample set, denoted as C_{avgH} and C_{avgL}, respectively, to obtain the CNN feature evaluation standard interval Cnorm, as shown in Formula 10.
4. Finally, calculate the final score $Grade_{CNN}$ by computing the distance between the probability value C_{test} of the sample under test and C_{avgH}, and then comparing it with Cnorm. If C_{test} is greater than or equal to C_{avgH}, the current indicator is marked as full marks; if C_{test} is less than C_{avgH}, the calculation formula is as shown in Formula 11.

$$C_{norm} = C_{avgH} - C_{avgL} \quad \text{Formula 10}$$

$$\text{Grade}_{\text{CNN}} = 1 - \frac{C_{avgH} - C_{test}}{C_{norm}} \quad \text{Formula 11}$$

The calculation process for the centroid feature score is as follows:

1. Input the test sample into the recognition model to obtain the classification result of the test sample;
2. Retrieve data based on the classification result to obtain the sample sets D_{good} and D_{bad};
3. Calculate the centroid coordinates of each sample using formulas 1–5;

4. Calculate the average centroid coordinates of the sample sets D_{good} and D_{bad}, denoted as Gavgg and Gavgb, respectively;
5. Calculate the centroid coordinates of the test sample using formulas 1–5, denoted as G_{test};
6. Substitute G_{test}, G_{avgg}, and G_{avgb} into formula 6 to calculate the centroid feature score Gradegravity. During calculation, if the cosine similarity between G_{test} and C_{avgH} reaches 100%, the current indicator is marked as full score; otherwise, the calculation is completed using formula 6.

The calculation process for the grid feature score is as follows:

1. Input the test sample into the recognition model to obtain the classification result of the test sample;
2. Retrieve data based on the classification result to obtain the sample sets D_{good} and D_{bad};
3. Calculate the pixel density matrix for each sample using formulas 7–9;
4. Calculate the average pixel density matrix for the sample sets D_{good} and D_{bad} and perform flattening operations, denoted as $Grid_s$ and $Grid_b$, respectively;
5. Calculate the pixel density matrix for the test sample using formulas 7–9 and perform flattening operations, denoted as $Grid_t$;
6. Substitute $Grid_t$, $Grid_s$, and $Grid_b$ into Formula 12 to calculate the grid feature score $Grade_{grid}$ for the sample. During calculation, if the cosine similarity between $Grid_t$ and $Grid_s$ reaches 100%, the current indicator is marked as full score; otherwise, the calculation is completed according to Formula 12.

$$Grade_{\text{grid}} = \left(1 - \left(1 - \frac{\text{Grid}_t \cdot \text{Grid}_s}{\|\text{Grid}_t\| \cdot \|\text{Grid}_s\|}\right) \Big/ \left(1 - \frac{\text{Grid}_s \cdot \text{Grid}_b}{\|\text{Grid}_s\| \cdot \|\text{Grid}_b\|}\right)\right) \times 100 \quad \text{Formula 12}$$

After calculating the three quantitative features, a fuzzy comprehensive evaluation is conducted. By incorporating fuzzy mathematics theory, a comprehensive evaluation of multiple indicators is achieved. Assuming the weight vector is W = (w1, w2, w3), where w1 + w2 + w3 = 1, where w1 corresponds to the weight of the "accuracy" indicator, w2 corresponds to the weight of the "balance" indicator, and w3 corresponds to the weight of the "standardization" indicator. The weight values for the three quantitative indicators are determined using a neural network model for calculation. The model is trained using the results of images with labeled scores to derive the weight distribution of the quantitative indicators, ensuring that the proportions of each indicator in the comprehensive evaluation align as closely as possible with actual evaluation requirements, thereby achieving intelligent evaluation.

4. Quantitative evaluation model experiment

4.1. Data preparation

The experimental dataset primarily originates from the publicly available CASIA-HWDB1.1 dataset and some custom-made data (with an additional 20 custom-made samples per category featuring better handwriting quality; hereinafter referred to as the dataset). CASIA-HWDB1.1 is a single-character handwritten Chinese character dataset released by the Institute of Automation, Chinese Academy of Sciences, containing 1,172,907 sample images of 3,755 commonly used Chinese characters written by 300 writers [8]. Among these, 60% are used as the training set, 20% as the validation set, and 20% as the test set. Since the sample sizes in the dataset vary, this study adopts a method of white border expansion combined with an affine transformation to standardize the sizes. This is because white border filling uses background pixels, which do not affect the recognition training of Chinese characters. Additionally, since each sample set contains 200–260 data samples, the data volume is relatively low

after division. Therefore, this study performs data augmentation on the original data, using an affine transformation combined with an elastic transformation to increase the number of samples.

4.2. Experimental steps

The experimental steps for the quantitative evaluation model are as follows:

1. Divide the preprocessed dataset into training, testing, and validation sets in a 6:2:2 ratio. The training set is used for iterative model training and parameter learning, the validation set is used to monitor the model training process and adjust hyperparameters, and the testing set is used to evaluate the final model performance.
2. Build a multi-feature fusion recognition module (Multi-model) based on the model structure diagram in Section 3.3, incorporating a BatchNormalization layer and an adjusted Inception submodule. Train the model, save it after it stabilizes, and add a hard voting mechanism to optimize the recognition results.
3. Quantify metric weight training. Train a neural network model to determine the weight vector, optimize the weight parameters using the backpropagation algorithm, and determine weight values that meet actual requirements.
4. Based on the recognition model, the "correctness" metric ($Grade_{CNN}$), "balance" metric ($Grade_{gravity}$), and "normativity" metric ($Grade_{grid}$) are calculated for each sample in the training set, and these metrics are used as input features for fuzzy comprehensive evaluation.
5. Three weighting methods are proposed, two of which use specified weight ratios (1:1:1 and 2:1:1 for $Grade_{CNN}$, Gradegravity, and $Grade_{grid}$, respectively), while the third uses an intelligent evaluation weight ratio calculated via the backpropagation algorithm (39:32:29). Comparative experiments are conducted for the three weighting allocation schemes.

4.3. Evaluation of effects and analysis of results

Based on the evaluation dimensions outlined in this paper, four main categories of image types were selected as experimental subjects: aesthetically pleasing handwriting images, handwriting position offset images, handwriting tilt images, and poor handwriting images. These four categories of images were designated as Class A, Class B, Class C, and Class D images, respectively, for ease of description in the subsequent text. During the experiment, testing was conducted under three weighting schemes (1:1:1, 2:1:1, Intelligent Evaluation Weighting 39:32:29) to assess the differences between the scores of each indicator, the comprehensive evaluation results, and the actual expert scores.

Class A images are samples with good writing quality (**Figure 3**). Taking two images as examples, one is a self-made image named A1 (average expert score of 87.08), and the other is an image from the CASIA-HWDB1.1 dataset named A2 (average expert score of 84.88). Both images perform excellently in terms of "correctness", "balance", and "standardization." A1's $Grade_{CNN}$ score is 86, $Grade_{gravity}$ score is 86, $Grade_{grid}$ score is 90, the comprehensive score under the 1:1:1 weighting scheme is 87.33, the comprehensive score under the 2:1:1 weighting scheme is 87, and the intelligent comprehensive score is 87.16. For A2, $Grade_{CNN}$ scored 75, $Grade_{gravity}$ scored 92, $Grade_{grid}$ scored 90, the 1:1:1 weighting scheme had a composite score of 85.67, the 2:1:1 weighting scheme had a composite score of 83, and the intelligent composite score was 84.79.

Figure 3. Beautifully written images (Class A images, left image is A1, right image is A2)

Category B images are samples with positional shifts (**Figure 4**). Taking two images as examples, one is a self-created image named B1 (with an average expert score of 80.88), and the other is an image from the CASIA-HWDB1.1 dataset named B2 (average expert score of 79.67). Both images perform excellently in terms of "correctness" and "standardization" features, but their "balance" performance is slightly weaker. B1's $Grade_{CNN}$ score is 83, $Grade_{gravity}$ score is 72, $Grade_{grid}$ score is 90, the comprehensive score under the 1:1:1 weighting scheme is 81.67, the comprehensive score under the 2:1:1 weighting scheme is 82, and the intelligent comprehensive score is 81.51. For B2, $Grade_{CNN}$ scored 77, $Grade_{gravity}$ scored 67, $Grade_{grid}$ scored 87, the 1:1:1 weighting scheme had a composite score of 77, the 2:1:1 weighting scheme had a composite score of 77, and the intelligent composite score was 76.7.

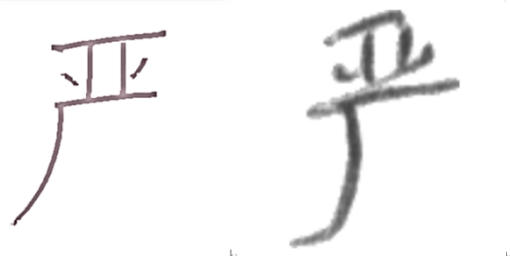

Figure 4. Writing offset images (Class B images, left image is B1, right image is B2)

Category C images are samples of slanted handwriting (**Figure 5**). Taking two images as examples, one is a self-created image named C1 (average expert score of 65.05), and the other is an image from the CASIA-HWDB1.1 dataset named C2 (average expert score of 61.10). Both images perform excellently in terms of "correctness", "balance", and "standardization." C1's $Grade_{CNN}$ score is 75, $Grade_{gravity}$ score is 37, $Grade_{grid}$ score is 82, the comprehensive score under the 1:1:1 weighting scheme is 64.67, the comprehensive score under the 2:1:1 weighting scheme is 67.25, and the intelligent comprehensive score is 64.87. C2's $Grade_{CNN}$ score is 68, $Grade_{gravity}$ score is 38, $Grade_{grid}$ score is 62.67, the comprehensive score for the 1:1:1 weighting scheme is 62.67, the comprehensive score for the 2:1:1 weighting scheme is 64, and the intelligent comprehensive score is 62.46.

Figure 5. Handwritten slanted images (Class C images, left image is C1, right image is C2)

Category D images (**Figure 6**) are test image samples with poor handwriting, named D1 (average expert score of 45.67). For D1, the $Grade_{CNN}$ score is 42, the $Grade_{gravity}$ score is 36, the $Grade_{grid}$ score is 65, the composite score for the 1:1:1 weighting scheme is 47.67, the composite score for the 2:1:1 weighting scheme is 46.25, and the intelligent composite score is 46.75.

Figure 6. Poorly written image (Class D image)

Based on experiments conducted using three weighting schemes, the experimental data are presented in the Weighting Comparison Experiment **Table 1**. Since the features selected in this paper have a certain degree of correlation in terms of aesthetic evaluation, the scoring results under the three weighting schemes show only minor differences in error. However, based on the overall experimental results, it can be observed that under the same experimental conditions, the comprehensive scores derived from the weights obtained through the backpropagation algorithm exhibit generally lower errors compared to the expected scores, thereby better aligning with the expected evaluation capabilities.

Table 1. Weight comparison experiment table

Experimental subjects	$Grade_{CN}$	$Grade_{gravity}$	$Grade_{grid}$	1:1:1	2:1:1	39:32:29	Expected score
A1	86	86	90	87.33	87	87.16	87.08
	error value			0.25	-0.08	0.08	
A2	75	92	90	85.67	83	84.79	84.88
	error value			0.79	-1.88	-0.09	
B1	83	72	90	81.67	82	81.51	80.88
	error value			0.79	1.12	0.63	
B2	77	67	87	77	77	76.7	79.67
	error value			-2.67	-2.67	-2.97	
C1	75	37	82	64.67	67.25	64.87	65.05
	error value			-0.38	2.2	-0.18	
C2	68	38	82	62.67	64	62.46	61.1
	error value			1.57	2.9	1.36	
D1	42	36	65	47.67	46.25	46.75	45.67
	error value			2.0	0.58	1.08	

5. Summary and outlook

This paper addresses the issues of subjectivity and the lack of unified quantitative standards in traditional Chinese character evaluation. It proposes a three-dimensional evaluation framework based on CNN features, center of gravity features, and grid features, integrating micro- and macro-level evaluation content. An improved multi-model convolutional neural network (combining VGG16 and GoogLeNet features) is used to extract stroke details, quantifying "correctness." Image moments are calculated to determine the center of gravity coordinates, measuring "balance" to assess whether Chinese characters are tilted or misaligned. A 3×3 grid division and pixel density comparison are employed to evaluate "standardization", ensuring proper structural integrity. This ultimately forms a closed-loop design of "recognition—quantification—fuzzy comprehensive evaluation." The recognition effect is optimized through a hard voting mechanism, and then a three-dimensional metric evaluation range is calculated based on the "best-worst sample set." Finally, intelligent weights (39:32:29) are trained through backpropagation. The experiment was validated using the CASIA-HWDB1.1 dataset and custom samples, evaluating four common types of images. The results showed that under dynamic weights, the average error between comprehensive scores and expert scores was generally lower than under fixed weights, with excellent performance across different sample types, such as calligraphic aesthetics and tilt. This effectively achieved the objectification and quantification of Chinese character evaluation. while a fixed weighting ratio of 1:1:1 is more suitable for scenarios where evaluation criteria have equal importance, and a weighted ratio of 2:1:1, which is more biased, is more suitable for evaluation scenarios with special requirements.

There are still some areas in this study that require further optimization. The dataset used in this experiment primarily consists of images from the CASIA-HWDB1.1 dataset, with a relatively small proportion of self-generated images. As a result, the model performs well in tasks evaluating images similar in style to the CASIA-HWDB1.1 dataset, but its performance is lower for self-generated images. In future research, the authors will further explore the following areas: first, expanding the dimensions of the samples to include those generated by different input devices (such as handwriting and smart device writing); second, adding an evaluation dimension for "fluency" in writing to expand the model's application scenarios; Third, further refine the dynamic weighting settings and introduce scenario-based weighting allocation functionality, enabling dynamic adjustment of evaluation metrics for different user groups, such as prioritizing "standardization" for children and "accuracy" for foreign learners.

Funding

Chinese Character Learning System Based on Multi-feature Fusion HCCR Intelligent Recognition and Evaluation Model, No.: KJQN202305604

Supported by the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant No. KJQN202305604)

Disclosure statement

The authors declare no conflict of interest.

References

[1] Xiao X, Li CC, 2022, Research progress on handwritten Chinese character evaluation methods. Computer Engineering and Application, 58(2): 27–42.

[2] Zhuang CB, Jin LW, 2005, An Intelligently Verified Algorithm for Correctness and Calligraphy of Online Handwritten Chinese Characters. Proceedings of the 12th National Annual Conference on Signal Processing (CCSP-2005).

[3] Han RF, An WH, Xun ED, et al., 2016, Real-time Grading Judgement for Stroke Quality in Chinese Character Handwriting. Journal of Computer Applications, 36(S1): 281–285.

[4] Yu K, Wu JQ, Zhuang YT, Calligraphic Character Retrieval Based on Skeleton Similarity. Journal of Computer-Aided Design & Computer Graphics, 21(6):746–751.

[5] Wu CZ, 2017, Research and Implementation of Evaluation System of Chinese Calligraphy Copy, thesis, South China University of Technology.

[6] Ge JM, 2016, Aesthetic Evaluation of Robot Calligraphy Based on Probability Distribution, thesis, Xiamen University.

[7] Zhuang ZM, 2019, Handwritten Chinese Character Recognition and Aesthetic Grading Based on Deep Learning, thesis, Beijing University of Posts and Telecommunications.

[8] Deng P, 2020, Research on Multi-feature Fusion Sample R, thesis, Recognition and Evaluation Model of Chinese Character Handwriting, thesis, Chongqing University. https://doi.org/10.27670/d.cnki.gcqdu.2020.001457