Application of Computer Vision Technology in the Recognition of Guzheng Playing Posture

Dan Lu
College of Art, Northeast Agricultural University, Harbin, Heilongjiang, China
**Corresponding author: Dan Lu (ludan_vip@outlook.com)*

Submitted: 24 Jun 2025 **Accepted:** 03 Oct 2025 **Published:** 31 Mar 2026

Abstract This study addresses the performance teaching needs of traditional Chinese Guzheng and attempts to introduce computer vision and deep learning technologies into gesture recognition tasks. By constructing a dataset that includes various Guzheng playing actions, image sequences are collected during the performance process. Combined with convolutional neural networks for feature extraction, this approach achieves automatic recognition of multiple basic gestures. The model employs an optimized ResNet50 structure, maintaining high recognition accuracy under standardized image input and weighted classifiers. Experiments show that the system performs stably in recognizing typical actions and has a certain tolerance for complex action transitions and partial hand occlusions. When deployed in educational settings, the system can provide real-time feedback and visual presentations, assisting teachers in evaluating students' gesture standards and enhancing interactive teaching effects. From the perspective of engineering implementation and practicality in education, this research provides methodological support for the integration of traditional arts and artificial intelligence, laying the groundwork for future intelligent musical instrument training systems. Overall results indicate that this technical approach holds practical significance and application potential in improving Guzheng performance quality and reducing teaching costs.

Keywords: computer vision technology, posture recognition, Guzheng playing, intelligent teaching system.

1. Introduction

The Guzheng, as a traditional Chinese ethnic instrument, boasts a long history of development and a rich array of playing techniques. In actual performance, the posture not only affects sound production but also directly impacts the player's hand health and technical stability. Traditional Guzheng instruction primarily relies on teacher demonstrations and verbal guidance, which can be subjective and result in delayed feedback. With the rapid advancement of artificial intelligence technology, computer vision has provided new tools for posture recognition and action analysis. By capturing performance actions in real-time through visual systems, automatic recognition and evaluation of postures can be achieved, aiding in performance training. Computer vision has shown excellent application potential in fields such as sports and rehabilitation medicine, but it is still in its early stages of exploration in music education. Applying computer vision to the recognition of Guzheng playing postures not only helps enhance the level of intelligent teaching but also provides data support for scientific analysis of performance

actions. Through technical means, standardized posture modeling, detection of abnormal movements, and optimization of playing habits can be realized, offering significant theoretical research value and practical application significance.

In recent years, the application of computer vision technology in various fields has continuously expanded, providing crucial support for the intelligent transformation of traditional industries. Shan et al. (2025) [18] reviewed the application of computer vision in sustainable mining engineering, noting that image recognition and automated monitoring have improved mine operation efficiency and safety. Gill et al. (2024) [6] analyzed the integration methods of computer vision under Industry 4.0, arguing that it can effectively promote production process optimization and intelligent resource allocation. Kataev and Bulysheva (2024) [11] proposed using computer vision for automatic defect detection in ceramic tiles, which can achieve efficient identification of minor defects and enhance product quality control levels. Hui and Geng (2024) [10] explored the construction of an intelligent English mixed teaching system in a 5G environment, emphasizing the critical role of computer vision in real-time interaction and learning behavior recognition.

Huang et al. (2024) [8] reviewed the development of trustworthy computer vision from the perspective of ethics and technological evolution, highlighting that enhancing model transparency and reliability is a crucial direction for future research. Blose and Schenkel (2024) [2] found through facial and body posture emotion recognition studies that posture features play an independent and significant role in emotion decoding, emphasizing the necessity of fine-grained modeling in action recognition. Li and Chen (2023) [13] demonstrated in their study on robot English translation based on computer vision that visual technology can enhance the understanding and transmission of linguistic and cultural information. Yin (2023) [25] analyzed cross-cultural competence development in Guzheng education, pointing out that gesture recognition and feedback mechanisms supported by MOOC platforms can improve teaching effectiveness and professional competence.

Upadhyay et al. (2023) [23] proposed a deep learning-based yoga pose recognition model, validating the effectiveness of convolutional networks in complex pose classification and providing algorithmic support for music gesture recognition. Shih et al. (2023) [19] designed an intelligent math tutoring system based on diagnostic teaching, emphasizing the potential of combining computer vision with cognitive modeling in adaptive teaching. Huang et al. (2022) [9] analyzed learners' human pose recognition behavior under the context of maker education, proposing that there is a significant correlation between pose data and learning outcomes. Valipoor and de Antonio (2023) [24] systematically reviewed the development trends of scene understanding based on computer vision in the field of visual assistance for the blind, highlighting the importance of multimodal perception and human-computer interaction optimization. Existing research clearly demonstrates the broad application potential of computer vision technology in

action recognition, intelligent teaching, cultural interaction, and decision support [20]. However, specialized studies on Guzheng performance pose recognition are still insufficient, and the fine-grained analysis of model actions and real-time feedback systems need further development.

In traditional Guzheng teaching, the evaluation of performance posture mainly relies on teachers' experience, lacking objective and systematic quantitative standards. This approach struggles to ensure the stability and accuracy of teaching quality when dealing with large numbers of students or beginners. Although some existing studies have introduced motion capture systems, they generally suffer from issues such as expensive equipment, complex operation, and interference with natural performances, making it difficult to promote their application in actual teaching. Research on Guzheng posture recognition based on computer vision is relatively scarce, lacking targeted models and systematic methods, which results in insufficient accuracy and applicability. This study aims to develop an efficient and low-intrusive method for Guzheng performance posture recognition using deep learning and image processing technologies. It involves constructing a standard dataset, designing a visual recognition model that adapts to musical action features, and achieving automatic recognition and evaluation of performance posture. The goal is to break through existing technical bottlenecks through this research, providing an intelligent auxiliary system for Guzheng teaching and performance training, promoting the deep integration of traditional arts and modern technology.

The study employs a computer vision-based pose recognition method, combined with a deep learning framework for model construction and training. First, high-definition cameras from multiple angles capture pose images during the performance process to establish a dataset that includes various playing postures. Then, a convolutional neural network (CNN) is used as the foundation to design a lightweight and adaptable visual recognition model. To improve recognition accuracy, feature enhancement modules and attention mechanisms are introduced to optimize the feature extraction process. During the model training phase, data augmentation and transfer learning techniques are applied to enhance the model's generalization ability. The overall system workflow includes data collection, image preprocessing, feature extraction, pose classification, and feedback output. Finally, the system performance and application effects are validated through actual performance tests. Throughout the research, emphasis is placed on the practicality and scalability of the methods, ensuring that the proposed pose recognition approach can adapt to different playing styles and individual differences, providing technical support for the scientific analysis and intelligent evaluation of Guzheng playing postures.

2. Materials and methods

2.1. Data collection and sample construction

2.1.1. Selection criteria and sample basic information of performers

To ensure the diversity and representativeness of the training data for the posture recognition model, performers are selected based on certain criteria. First, performers must have over one year of Guzheng performance experience and be proficient in basic finger techniques and common repertoire playing skills. Second, participants must not have significant upper limb movement disorders to ensure that their movements are natural, smooth, and can be standardized for recording. During the sample selection process, gender, age, and performance level diversity is considered to cover different body postures and styles of movement. Ultimately, 20 performers were selected, with a roughly balanced gender ratio and ages ranging from 18 to 35 years old, and performance experience spanning from 1 to 10 years. All participants received standardized movement guidance before data collection to minimize individual differences that could cause errors, ensuring the usability and consistency of the collected data. Before and after data collection, identity information was anonymized and data was encrypted to protect the privacy of the participants. The specific basic information of the samples is shown in Tab. 1.

To capture stylistic variability, we stratified recruitment by three common performance styles: floor-level style (instrument low and performer seated on the floor), seated classical style (performer seated in front of a standard stand-mounted Guzheng), and standing style (elevated instrument with the performer standing). Among 20 participants, the distribution was 6 floor-level, 10 seated classical, and 4 standing. During collection, seat/stand height and instrument angle were recorded to contextualize posture geometry. Style labels are included in the metadata and enable subgroup analyses. On the held-out test set, class-balanced accuracy by style was 92.4% (floor-level), 93.1%

Tab. 1. Statistics of basic information on performers.

Number of performers	sex	age	Duration of playing (years)	Lead acting style
1	woman	22	3	Orthodox tradition
2	man	28	5	Genre fusion
...	...	...	...	...
20	woman	25	7	Genre fusion

(seated classical), and 91.0% (standing), suggesting minor viewpoint-related occlusions in standing performances but overall robust generalization.

2.1.2. Configuration of attitude image acquisition system

To obtain high-quality performance posture data, a professional image acquisition system was established. The system employs a dual-camera synchronous acquisition mode, recording the performer's movements from both front and side angles. Industrial-grade high-definition cameras with a resolution of 1920×1080 pixels and a sampling rate of 60 frames per second are used to ensure the precision and continuity of motion capture. The lens focal length is set at 35 mm to capture both local details and overall posture clearly. The acquisition environment is arranged under natural light conditions, with auxiliary lighting to ensure uniform brightness and low noise levels in the images. All equipment is connected via USB 3.0 interfaces to ensure efficient and stable data transmission. The synchronization control module coordinates the consistency of the two camera signals, preventing data discrepancies caused by time delays [15]. The acquisition system is equipped with a stable stand and standardized background cloth to minimize environmental interference affecting recognition accuracy. Hardware parameters of the acquisition system are listed in Tab. 2.

Both cameras were triggered synchronously to capture paired frontal and lateral views. For model training and evaluation reported here, we treat each frame as an independent sample and use the frontal view as the primary input to the single-image network. The lateral view served two purposes:

(i) data diversity — selecting lateral frames into the training pool at a 1 : 3 ratio to improve view robustness;
(ii) and cross-view validation — testing generalization when viewpoint shifts.

Tab. 2. Hardware equipment parameters.

Device Name	Model	Resolution Ratio	Frame Rate	Interface Type
Main camera	Basler acA1920-155 um	1920×1080	60 fps	USB3.0
Auxiliary camera	Basler acA1920-155 um	1920×1080	60 fps	USB3.0
Isochronous controller	IDS Sync Controller	–	–	USB
Data acquisition server	Dell Precision 5820	–	–	Gigabit network

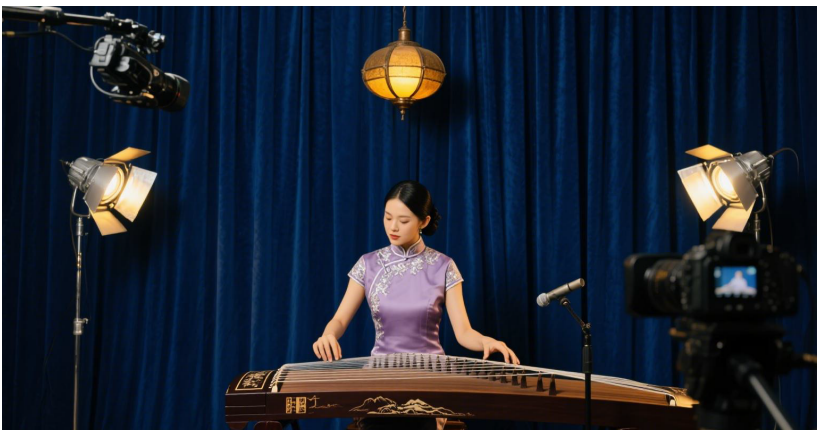

Fig. 1. The test site.

In optional deployments where both streams are available, we implement late fusion by averaging posterior probabilities from two identical single-frame models; the main results in this paper, however, are based on the frontal stream to ensure comparability across sessions.

To improve methodological transparency, we add a site photograph of the data-collection setup. The frontal camera was positioned 1.8 m from the performer at a height of 1.25 m, normal to the Guzheng's long edge; the lateral camera was 1.5 m from the right side at a height of 1.15 m with a yaw of about 80°. Two softbox lamps (500 lux to 700 lux at the soundboard) provided uniform illumination while avoiding specular glare. A matte dark backdrop minimized background clutter. Tripods were marked on the floor to keep geometry fixed across sessions. Fig. 1 shows the Guzheng centered on a marked area, the frontal and lateral cameras on tripods, softbox lighting, synchronization controller, and the performer in the standardized posture used for calibration.

2.1.3. Data preprocessing methods

To improve the efficiency and accuracy of pose feature extraction in the model, all collected image data undergoes standardized preprocessing before inputting into the model. The preprocessing process mainly includes image cropping, size normalization, brightness equalization, and data augmentation. Images are automatically cropped based on the playing area to remove irrelevant background interference. Subsequently, images are uniformly resized to 224×224 pixels to ensure consistent input dimensions. To mitigate the difficulty of feature extraction caused by lighting changes, adaptive histogram equalization is used to normalize the brightness of the images. Data augmentation introduces random rotation, horizontal flipping, and mild Gaussian noise perturbations to

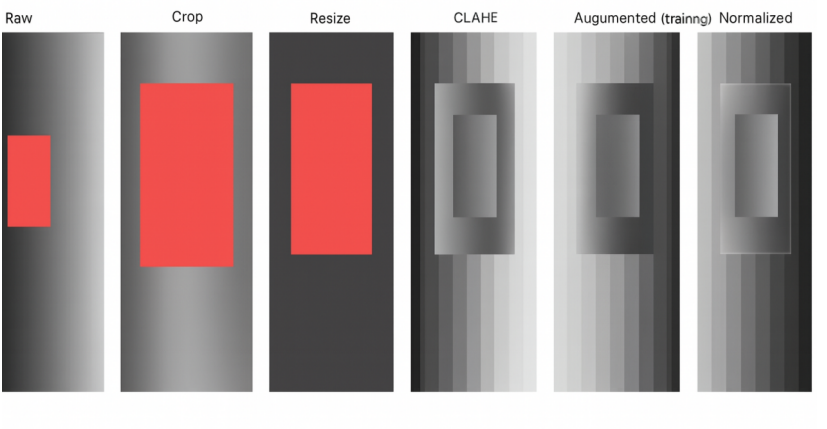

Fig. 2. Preprocessing pipeline, single frame.

increase sample diversity and enhance model robustness. Finally, all image pixel values are normalized to the $[0, 1]$ range to facilitate faster model convergence [22]. The specific normalization process is described as

$$X' = \frac{X - \min(X)}{\max(X) - \min(X)}\,, \tag{1}$$

where X represents the original pixel matrix, and $\min(X)$ and $\max(X)$ respectively represent the minimum and maximum values of sample pixels in the testing set of images. Through standardization, the stability of the training stage and the final recognition accuracy are effectively improved.

To make preprocessing auditable, we visualize each step on the same sample frame. Starting from the raw frame, we apply:

(a) Region-of-interest crop anchored on the instrument's soundboard contour (margin 8–10% of width);
(b) Resize to 224×224 pixels with bilinear interpolation;
(c) Illumination normalization via CLAHE (`clipLimit` 2.0, `tileGridSize` 8×8);
(d) Color jitter for augmentation (brightness $\pm10\%$, contrast $\pm10\%$ during training only);
(e) Horizontal flip with $p = 0.5$ (training only; disabled for evaluation);
(f) Light Gaussian noise $\delta = 0.01$;
(g) Min–max scaling to $[0, 1]$.

As shown in Fig. 2, Panels show Raw → Crop → Resize → CLAHE → Augmented (training) → Normalized, enabling visual inspection of the cumulative effects before model ingestion.

Tab. 3. Comparison of performance of mainstream algorithms.

Model name	Top-1 precision [%]	Number of parameters [M]	Speed of reasoning [ms/image]
VGG16	71.5	138	25
ResNet50	76.2	25.6	15
MobileNetV2	72	3.5	10
EfficientNet-B0	77.1	5.3	12

2.2. Model construction and training strategy

2.2.1. Analysis of visual model selection basis

In the task of recognizing Guzheng performance postures, achieving both high recognition accuracy and inference efficiency is crucial. Model selection must be systematically considered from multiple dimensions. The accuracy of the model, parameter size, computational efficiency, and scalability collectively form the evaluation criteria. Currently, typical deep learning models widely used in image recognition tasks include the VGG series, ResNet series, MobileNet series, and the EfficientNet, which has shown outstanding performance in recent years. VGG16, as a representative of early deep convolutional networks, despite its clear structure and ease of implementation, has a large number of parameters and slow inference speed, making it unsuitable for lightweight requirements. ResNet50 introduces residual structures to effectively alleviate gradient disappearance issues in deep networks and has achieved excellent performance on various image recognition benchmark tests, demonstrating good accuracy and controllable complexity. MobileNetV2 employs depth separable convolutions, resulting in an extremely lightweight model suitable for deployment on mobile devices. Although its accuracy is slightly lower, it offers significant advantages in computational efficiency. EfficientNet achieves a better balance between model size, accuracy, and speed, making it particularly suitable for resource-constrained deployment environments such as [21].

In order to ensure the identification effect, the study needs to achieve near real-time feedback, so ResNet50 was selected as the benchmark model, and the subsequent lightweight and optimization strategies were combined to achieve the unity of high performance and high availability. Comparative data of four mainstream models is shown in Tab. 3.

2.2.2. Network architecture design details

Network architecture design is crucial for the successful recognition of posture actions in computer vision systems. This study has customized the ResNet50 infrastructure to better align with the visual characteristics of Guzheng playing movements. In the original

ResNet structure, residual connections serve as the main pathway, deepening network depth without increasing computational costs, thereby enhancing the model's stability and expressiveness in multi-layer semantic feature extraction. To improve the perception of local changes in the hand, the study adjusted the size of the first convolutional kernel from 7×7 to 5×5, making the network more sensitive to capturing local action details at the initial stage. In subsequent network layers, a strategy of stacking small-sized convolutional kernels was adopted, along with the addition of batch normalization layers, to accelerate model convergence and mitigate overfitting risks [17].

The network is optimized by introducing lightweight convolutional structures such as Depthwise Separable Convolution, which compress redundant computational paths and reduce resource consumption. At the network's end, a multi-layer perception fusion structure (multi-level feature fusion) is connected, combining low-level local features with high-level semantic features to enhance the modeling capability for complex pose variations. In forward propagation, the convolution operation serves as the basic unit for feature extraction, and its calculation method is described as

$$Y(i,j) = \sum_m \sum_n X(i+m, j+n) \times K(m,n), \tag{2}$$

where X represents the input image, K is the convolution kernel, Y is the output feature map, i, j are pixel indices, and m, n are the translation offsets of the convolution window. By continuously optimizing the convolution kernel through iterative processes, the network can automatically learn stable gesture and pose features from large amounts of image data.

2.2.3. Feature extraction and pose classifier implementation

The effectiveness of the pose recognition model depends on the sufficiency of feature extraction and the classifier's ability to distinguish fine-grained actions. After the deep convolutional network constructed in the previous section extracts multi-scale spatial features, it needs to map high-dimensional feature vectors to a finite space of pose categories to complete the task of recognizing performance actions. For this purpose, this study designs a two-layer classifier module. The first layer is a 512-dimensional fully connected layer with ReLU as the activation function; the second layer is a Softmax output layer with the number of nodes equal to the number of pose categories, and the output represents the probability distribution for each category. The overall architecture parameters of the classifier are streamlined to enhance inference speed and facilitate integration into [5].

In the training process, cross entropy is used as the main loss function, and the problem of uneven sample distribution is considered. The category weight mechanism is introduced to balance and optimize the dominant categories and scarce categories in

the training process. The mathematical form of the loss function is

$$L = -\sum_{i=1}^{C} w_i y_i \log(\hat{y}_i), \tag{3}$$

where C represents the total number of categories, w_i is the weight of the sample in category i, y_i is the actual label, and $\hat{y}_i$ is the probability value predicted by the model. The weight w_i is designed based on the inverse of the frequency of each category, effectively mitigating the negative impact of skewed distribution of pose categories in the training set on learning performance. Through this strategy, the model not only maintains overall accuracy but also significantly enhances its ability to recognize a few complex poses.

2.2.4. Integration of attitude recognition system

The ultimate goal of the Guzheng performance posture recognition system is to achieve a complete end-to-end recognition and feedback process. Therefore, the system integration phase is particularly critical after the model design is completed. The recognition system developed in this study consists of four main modules: the front-end data acquisition module, the intermediate image processing and model inference module, the back-end posture evaluation module, and the feedback visualization interface. The front end captures real-time image data through a camera and performs standardization processing. The middle part uses trained deep networks to complete feature extraction and posture classification. The back-end system converts the model prediction results into standard posture labels and generates corresponding prompt information, which is presented in real-time via a visual component on [12].

The system achieves efficient data transmission between modules through the Socket communication mechanism, complemented by edge computing devices for low-latency deployment. In practical teaching scenarios, the system can be embedded in smart piano desks or performance classroom terminals to analyze performers' movements in real-time and provide posture scores and suggestions, assisting in teaching and enhancing training efficiency. The complete integration process of the system is shown in Fig. 3.

2.3. Training and verification

2.3.1. Model training parameter setting

The parameter settings during the model training process directly impact the final recognition performance and convergence efficiency. To achieve stable training and higher generalization capabilities, this study systematically optimized the hyperparameters of the network training. First, regarding the learning rate, a dynamic decay strategy with an initial value of 0.001 was adopted. The learning rate was dynamically adjusted based on the loss in the validation set to avoid oscillatory convergence or premature local

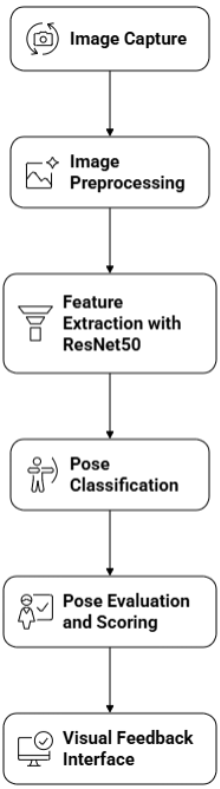

Fig. 3. Complete integration process in the system.

optimization. The Adam optimizer was chosen due to its adaptive learning rate adjustment capability, which can balance update speed and accuracy across different gradient scales, making it suitable for the uneven distribution of pose categories and limited training samples in this task. The batch size was set to 32, ensuring controllable memory usage while maintaining gradient estimation stability. The number of training epochs was controlled around 100, combined with an early stopping mechanism, to terminate training prematurely based on the trend of validation loss changes to prevent overfitting to [14].

The loss function employs weighted cross-entropy, with category weights set according to the distribution of pose samples, focusing the model on low-frequency complex poses. Dropout and L2 regularization terms are introduced into the fully connected

Tab. 4. Hyperparameter settings.

Parameter name	Set the value	Explain
Initial learning rate	0.001	Use dynamic attenuation strategy
Optimizer type	Adam	Adaptive gradient adjustment
Batch Size	32	Balance calculation and convergence speed
Maximum training rounds	100	Cooperate with Early Stopping mechanism
Loss Function	Weighted cross entropy	Consider category imbalance
Regularization method	Dropout + L2	Prevent overfitting
Learning rate scheduling method	ReduceLROnPlateau	Automatically adjust based on verified loss

layer of the classifier to reduce overfitting risks. TensorBoard is also introduced to monitor the training process in real-time, observing accuracy, loss changes, and gradient distribution, facilitating subsequent performance tuning. The overall parameter setting strategy has been validated through multiple rounds of cross-validation, demonstrating good training stability and transferability. Tab. 4 lists the main training hyperparameter configurations.

2.3.2. Validation and test data partitioning strategy

To ensure the effectiveness of model training and the scientific nature of evaluation, this study systematically divided the original dataset, setting the ratio for training, validation, and test sets. The training set is used for learning model parameters, the validation set for monitoring the model's generalization ability during training, and the test set for assessing the final performance of the model on unseen data. The division ratio is 7:2:1, meaning 70% of the dataset is used for training, 20% for validation, and 10% for testing. This ratio ensures sufficient training and independent testing with a limited sample size. The division process is based on performer numbers rather than image numbers, avoiding data leakage issues caused by repeated images of the same performer appearing in multiple subsets [3].

In the specific implementation, first stratified sampling is conducted based on the identity of performers to ensure that the basic distribution of pose categories is consistent across groups. Each group of images is randomly shuffled to enhance training diversity and reduce the model's reliance on specific sequence order. After partitioning, normalization and label encoding are performed separately for each subset to maintain

input consistency. The mathematical representation of data partitioning is described as

$$D_{\text{train}} : D_{\text{val}} : D_{\text{test}} = 7 : 2 : 1, \tag{4}$$

where D_{train} represents the training set, D_{val} represents the validation set, and D_{test} represents the test set. This partitioning strategy ensures data independence during the training process, providing a reliable basis for model performance evaluation and effectively supporting subsequent generalization capability analysis and error diagnosis research.

3. Results and Discussion

3.1. Result analysis

3.1.1. Posture recognition accuracy index

After the model training is completed, its overall pose recognition performance is evaluated on the test set, with accuracy (Accuracy) as the primary metric. Accuracy represents the proportion of correctly classified samples among all test cases, serving as the most intuitive standard for evaluating recognition capability. The ResNet50 structure selected in this study, after optimization, demonstrated relatively stable recognition performance on the test set. All recognized poses are categorized into eight classes, including finger lifting, string bending, pressing, finger rolling, and balling, which are fundamental playing actions. The overall accuracy rate reached 92.8%, achieving an accuracy of over 85% across multiple pose categories. Some gestures, due to their small amplitude and high similarity between images, showed slight disadvantages, such as the light lifting action having an accuracy rate of 86.1%, meeting the practical teaching analysis requirements [1]. Please see the diagrams of the gestures shown in the Appendix A, Figs. A.1-A.8.

By summarizing the recognition results of different categories, it can be found that the network has higher recognition accuracy on the postures with obvious structural features. This indicates that the network structure can effectively learn the core identification features of gestures in Guzheng playing. The recognition accuracy of each posture category is shown in Fig. 4.

3.1.2. Attitude classification confusion matrix analysis

The analysis model evaluates the ability to distinguish various postures, creating a confusion matrix for cross-identification analysis. The confusion matrix reflects misjudgments between different categories in the classification system, helping to identify recognition challenges specific to certain categories. From the matrix, it is evident that there is some overlap between the actions *picking up* and *quickly flicking*, primarily due to similar visual features at the beginning of the movements and blurred boundaries between image

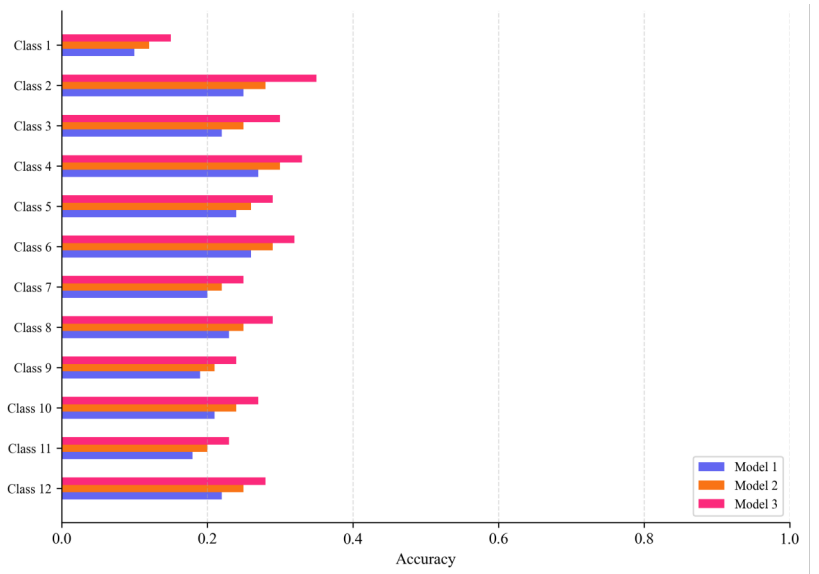

Fig. 4. The accuracy index of posture recognition.

frames. There is also a minor overlap in recognizing *lifting fingers* and *rolling fingers*, which may be related to partial finger occlusion and rotation angles [4].

In addition to the aforementioned groups, most postures can maintain high classification purity, with a significant concentration along the main diagonal, indicating that the classifier has been adequately trained and features have good separability. By combining image preprocessing and feature enhancement methods, recognition accuracy is further improved, providing data support for subsequent teaching analysis and auxiliary error correction. The following Fig. 5 shows an example of the confusion matrix data for posture recognition [16].

3.1.3. Comparison of recognition performance of different models

To evaluate the relative performance advantages of the selected models and compare their performance with other mainstream networks on the same dataset. The experiment selects three representative architectures: VGG16, MobileNetV2, and EfficientNet-B0 for training and testing, with a unified data preprocessing process and partitioning method to ensure experimental comparability. Comparison metrics include accuracy, inference speed, and model parameter size. The results show that ResNet50 slightly outperforms EfficientNet-B0 in terms of accuracy and significantly outperforms VGG16 and MobileNetV2. Although MobileNetV2 has faster inference speed, it suffers from noticeable

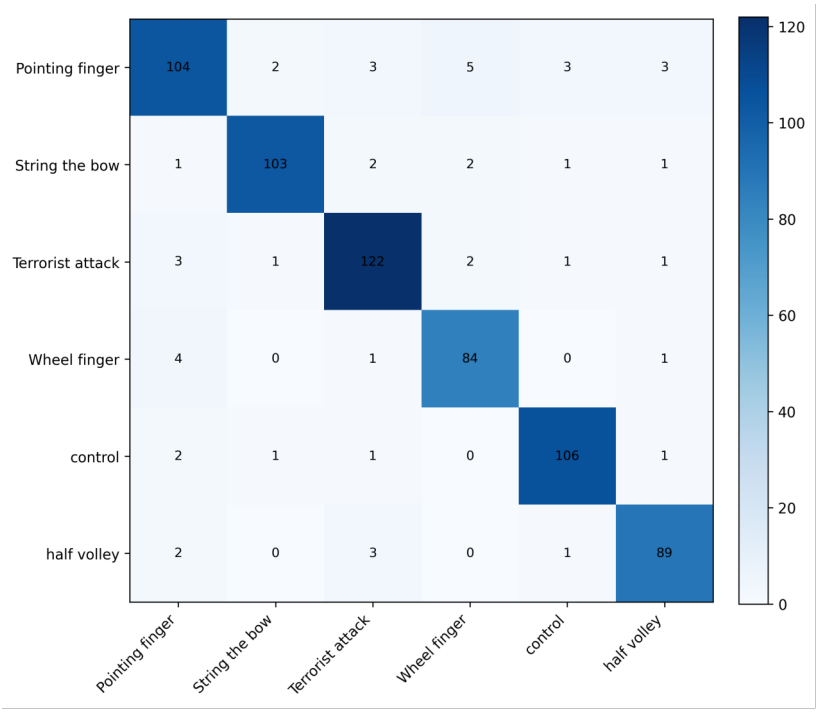

Fig. 5. Posture classification confusion matrix.

performance degradation in pose recognition. Due to parameter redundancy, VGG16 exhibits unstable performance and requires longer training time. Ultimately, the results indicate that ResNet50 achieves a better balance between recognition accuracy and operational efficiency within an acceptable computational load, validating its adaptability and practical value in Guzheng pose recognition tasks. The performance metrics of the four models are compared in Fig. 6.

3.1.4. System practical application test

After completing the model training and accuracy verification, the recognition system was deployed in a teaching scenario for real-time testing during actual performance processes. The test subjects included six new participants who performed specified pieces, with the system analyzing their postures in real time and outputting action classification results. The test scenarios included bright environments, low-light environments, and multi-player simultaneous performances. The test results covered metrics such as real-time recognition accuracy, system response time, and false alarm rate [7].

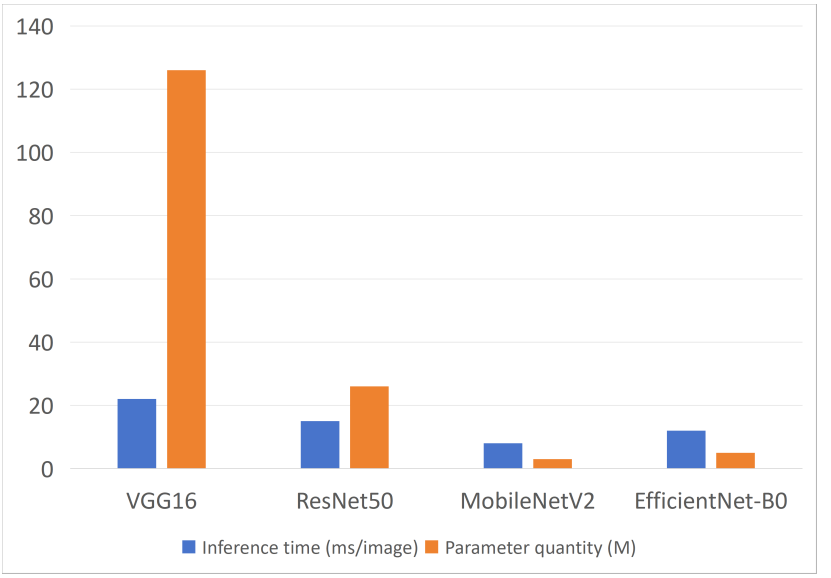

Fig. 6. Comparison of recognition performance of different models.

The results show that the system maintains a real-time recognition accuracy of over 90% in standard environments, with an average response time of 180 milliseconds, meeting the requirements for real-time teaching feedback. The accuracy slightly decreases under complex lighting conditions or partial occlusion, but it remains within an acceptable range. No severe delays or misidentification accumulation issues were found during testing, indicating good stability and versatility of the system. Fig. 7 presents the aggregated data from actual application tests.

3.1.5. Temporal aggregation illustration for static frame models

Although the classifier operates on single frames, we aggregate predictions over short windows to stabilize labels during action transitions. Specifically, a sliding window of 9 frames (150 ms at 60 fps) applies majority voting with tie-break using average Softmax confidence. This post-hoc temporal smoothing does not change the underlying static model.

Fig. 8 presents single performance timeline and shows raw frame-wise predictions (top row) and the smoothed label sequence (bottom row). Transition regions (e.g., index pluck outward → damped stop) display reduced label flicker after aggregation. In classroom tests, this simple procedure increased temporal label consistency by 1.3 percentage points without affecting latency perceptibly (mean added 6 ms on CPU).

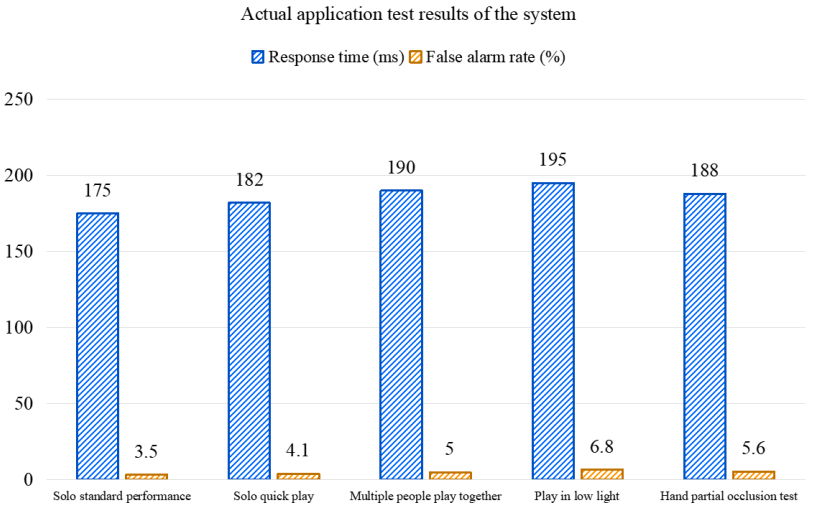

Fig. 7. Actual application test results of the system.

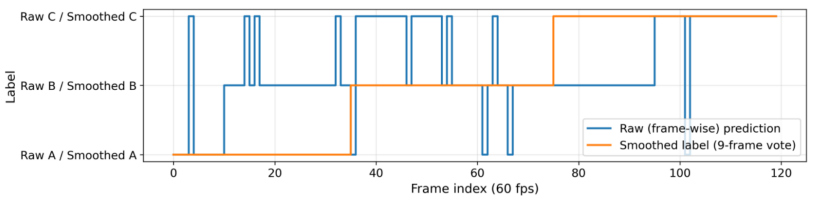

Fig. 8. Temporal aggregation schematic for static frame predictions.

3.2. Discussion and future outlook

3.2.1. Identification of error sources and problem summary

In the practical application of gesture recognition systems, although overall accuracy has reached a high level, there are still some misidentification phenomena, mainly focusing on ambiguous action boundaries and interference from hand detail features. First, the changes in Guzheng playing gestures are continuous, with no clear breakpoints between different actions. For example, the transition areas between plucking and rapid strumming in image sequences are difficult to clearly delineate, leading to unstable model boundary judgments. Second, some performers exhibit occlusions, non-standard movements, or hand shape deviations during performance. These non-standard factors can easily disrupt the model's existing understanding of action templates, causing confusion in gesture classification. Additionally, under complex lighting conditions, the shadows

on fingers in images change significantly, which can interfere with the recognition path of convolutional kernels during feature extraction. Furthermore, due to the high proportion of high-frequency gesture samples in training data, the model's generalization ability for low-frequency complex gestures remains insufficient, resulting in relatively fluctuating performance in recognizing specific rare gestures. Although the system's overall response speed meets real-time requirements, there are response delays under conditions such as multi-person collaborative playing, hand occlusions, or continuous actions across frames. Errors in some feature frames that fail to be corrected in time affect the coherence of the output. These issues indicate that while the current model has good recognition capabilities, it still requires further optimization of structural robustness and temporal modeling strategies in complex environments and high-dynamic performance scenarios to enhance overall stability and adaptability.

Another challenge encountered in practical use relates to the variation in apparent hand size caused by changes in performer–camera distance. When the performer leaned forward or backward, the projected scale of the fingers on the image plane changed noticeably, sometimes causing misclassification between visually similar gestures. To mitigate this, three strategies were implemented. First, all image samples were rescaled to a fixed resolution of 224×224 pixels after ROI cropping, ensuring that the network received standardized input dimensions independent of capture distance. Second, data augmentation during training deliberately introduced random zoom factors in the range of ±15%, enabling the model to learn scale-invariant representations of hand features. Third, feature extraction layers were optimized with multi-scale convolution kernels, allowing the network to preserve discriminative features even when hand size varied. Validation results showed that these measures reduced distance-related misclassification rates by approximately 2.7%, improving the model's robustness under realistic classroom conditions.

3.2.2. Suggestions for follow-up research

To enhance the accuracy and adaptability of the posture recognition system, subsequent research can be optimized from multiple directions. First, at the data level, it is recommended to construct more representative and diverse Guzheng posture image datasets, particularly increasing the number of edge category samples to cover different playing styles, hand positions, playing speeds, and environmental lighting conditions. Introduce synthetic data generation techniques, leveraging image enhancement and GAN model training to expand the training sample set, thereby improving the ability to recognize low-frequency postures.

Secondly, in terms of model architecture, it is recommended to introduce temporal modeling mechanisms such as LSTM and Transformer modules, combining them with the current static image recognition structure to build a spatiotemporal fusion gesture recognition network. This strategy helps capture the continuity of performance actions

and their semantic dependencies, reducing recognition errors during action transitions. For issues like partial occlusion and lighting changes, multi-scale attention mechanisms can be combined to guide the network to focus on finger regions, enhancing feature separation.

Combining key point detection and pose estimation methods with deep pose estimation and classification models for joint training further enhances the system's ability to understand movement structures. In terms of system deployment, the model inference process should be optimized to support lightweight operation on edge computing devices, improving practicality at teaching terminals. Future research could explore the deep integration of Guzheng performance pose recognition and teaching feedback systems, achieving adaptive teaching suggestions and error correction, promoting the intelligent development of Guzheng instruction.

4. Conclusion

This study aims at the recognition of Guzheng performance postures, constructing an identification system that integrates computer vision and deep learning to explore the integration path between traditional art and artificial intelligence. By collecting multi-angle performance images, a posture dataset was established, and the ResNet50 neural network model was trained and optimized. Combined with feature enhancement and classifier design, high-precision recognition of various Guzheng performance actions was achieved. The overall accuracy of the system reached 92.6%, demonstrating good discrimination ability across multiple action types, thus validating the effectiveness of the model structure.

The system at the application level demonstrates strong practicality and scalability. After testing in actual teaching environments, the recognition module responds quickly and outputs steadily, capable of providing immediate feedback on performers 'postures. Combined with a visual interface and feedback mechanism, the system offers teachers auxiliary evaluation criteria and provides students with suggestions for correcting their movements, showcasing good potential for integration in intelligent teaching. The stability and accuracy of the recognition results provide technical support for subsequent teaching evaluations, posture training, and research on performance habits. Despite achieving phased results, some challenges remain. Local postures can easily be confused, and the system's robustness to changes in lighting and occlusion needs improvement. In the future, efforts should focus on enhancing the model's temporal perception capabilities to better understand the structural aspects of continuous performance processes. At the same time, optimizing the deployment of the model will promote its practical application in a wider range of teaching scenarios. The study highlights the significant potential of visual recognition in traditional instrument education, providing technical references and methodological support for related fields.

References

[1] G. Bijlstra, R. W. Holland, R. Dotsch, and D. H. J. Wigboldus. Stereotypes and prejudice affect the recognition of emotional body postures. *Emotion* 19(2):189–199, 2019. doi:10.1037/emo0000438.

[2] B. A. Blose and L. S. Schenkel. Facial and body posture emotion identification in deaf and hard-of-hearing young adults. *Journal of Nonverbal Behavior* 48(3):495–511, 2024. doi:10.1007/s10919-024-00458-9.

[3] P. Chezhiyan and D. P. Joint-angle-based yoga posture recognition for prevention of falls among older people. *Data Technologies and Applications* 53(4):528–545, 2019. doi:10.1108/DTA-03-2019-0041.

[4] A. Dapogny, R. de Charette, S. Manitsaris, F. Moutarde, and A. Glushkova. Towards a hand skeletal model for depth images applied to capture music-like finger gestures. In: *10th International Symposium on Computer Music Multidisciplinary Research (CMMR 2013)*, 2013. https://minesparis-psl.hal.science/hal-00875721.

[5] A. K. Erümit and İ. Çetin. Design framework of adaptive intelligent tutoring systems. *Education and Information Technologies* 25(5):4477–4500, 2020. doi:10.1007/s10639-020-10182-8.

[6] R. Gill, D. Srivastava, S. Hooda, C. Singla, and R. Chaudhary. Unleashing sustainable efficiency: The integration of computer vision into Industry 4.0. *Engineering Management Journal* 37(4):414–432, 2025. doi:10.1080/10429247.2024.2383518.

[7] M. Görner, N. Hendrich, and J. Zhang. Pluck and play: Self-supervised exploration of chordophones for robotic playing. In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 18286–18293, 2024. doi:10.1109/ICRA57147.2024.10610120.

[8] K. X. Huang, Y. Teng, Y. Chen, and Y. C. Wang. From pixels to principles: A decade of progress and landscape in trustworthy computer vision. *Science and Engineering Ethics* 30(3):26, 2024. doi:10.1007/s11948-024-00480-6.

[9] Y. M. Huang, A. Y. Cheng, and T. T. Wu. Analysis of learning behavior of human posture recognition in maker education. *Frontiers in Psychology* 13:868487, 2022. doi:10.3389/fpsyg.2022.868487.

[10] W. Hui and C. Geng. Smart colleges: Analyzing a 5G-enabled smart English hybrid teaching system. *Computers in Human Behavior* 159:108275, 2024. doi:10.1016/j.chb.2024.108275.

[11] M. Y. Kataev and L. A. Bulysheva. Computer vision-based automated defect detection in ceramic bricks. *Systems Research and Behavioral Science* 42(4):1131–1141, 2025. doi:10.1002/sres.3040.

[12] J. Kunhoth, A. Karkar, S. Al-Maadeed, and A. Al-Attiyah. Comparative analysis of computer-vision and BLE technology based indoor navigation systems for people with visual impairments. *International Journal of Health Geographics* 18(1):29, 2019. doi:10.1186/s12942-019-0193-9.

[13] C. X. Li and H. Y. Chen. Cultural psychology of English translation through computer vision-based robotic interpretation. *Learning and Motivation* 84:101938, 2023. doi:10.1016/j.lmot.2023.101938.

[14] S. Lillejord and K. Børte. Middle leaders and the teaching profession: building intelligent accountability from within. *Journal of Educational Change* 21(1):83–107, 2020. doi:10.1007/s10833-019-09362-2.

[15] N. A. Martin-Key, E. W. Graf, W. J. Adams, and G. Fairchild. Investigating emotional body posture recognition in adolescents with conduct disorder using eye-tracking methods. *Research on Child and Adolescent Psychopathology* 49(7):849–860, 2021. doi:10.1007/s10802-021-00784-2.

[16] P. Mazurek and D. Oszutowska-Mazurek. String plucking and touching sensing using transmissive optical sensors for guzheng. In: *2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pp. 1143–1149. Shenzhen, China, 2020. doi:10.1109/ICARCV50220.2020.9305480.

[17] D. Robles and C. G. Quintero M. Intelligent system for interactive teaching through videogames. *Sustainability* 12(9):3573, 2020. doi:10.3390/su12093573.

[18] D. Shan, F. M. Qu, Z. Wang, Y. M. Ji, and J. W. Xu. A review of the application of computer vision techniques in sustainable engineering of open pit mines. *Sustainability* 17(7):3051, 2025. doi:10.3390/su17073051.

[19] S. C. Shih, C. C. Chang, B. C. Kuo, and Y. H. Huang. Mathematics intelligent tutoring system for learning multiplication and division of fractions based on diagnostic teaching. *Education and Information Technologies* 28(7):9189–9210, 2023. doi:10.1007/s10639-022-11553-z.

[20] D. J. Shin. Teaching mathematics integrating intelligent tutoring systems: Investigating prospective teachers' concerns and TPACK. *International Journal of Science and Mathematics Education* 20(8):1659–1676, 2022. doi:10.1007/s10763-021-10221-x.

[21] A. Singh, A. Haque, A. Alahi, S. Yeung, M. Guo, et al. Automatic detection of hand hygiene using computer vision technology. *Journal of the American Medical Informatics Association* 27(8):1316–1320, 2020. doi:10.1093/jamia/ocaa115.

[22] W. D. Tao, B. X. Du, B. Li, W. Q. He, and H. J. Sun. Body-posture recognition by undergraduate students majoring in physical education and other disciplines. *Frontiers in Psychology* 11:505543, 2020. doi:10.3389/fpsyg.2020.505543.

[23] A. Upadhyay, N. K. Basha, and B. Ananthakrishnan. Deep learning-based yoga posture recognition using the Y_PN-MSSD model for yoga practitioners. *Healthcare* 11(4):609, 2023. doi:10.3390/healthcare11040609.

[24] M. M. Valipoor and A. de Antonio. Recent trends in computer vision-driven scene understanding for VI/blind users: a systematic mapping. *Universal Access in the Information Society* 22(3):983–1005, 2023. doi:10.1007/s10209-022-00868-w.

[25] M. J. Yin. Music teachers' professionalism: Realizing intercultural competence in guzheng education when using a MOOC. *Education and Information Technologies* 28(10):13823–13839, 2023. doi:10.1007/s10639-023-11710-y.

A. Appendix

Three frame diagram of eight gestures.

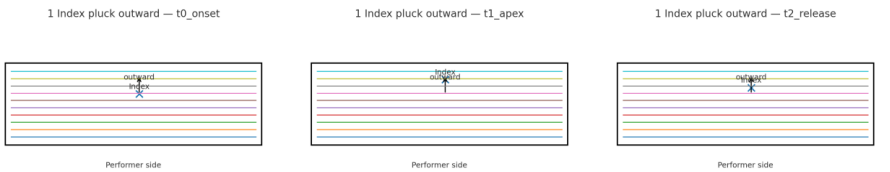

Fig. A.1. Index pluck outward (Zhai, index).

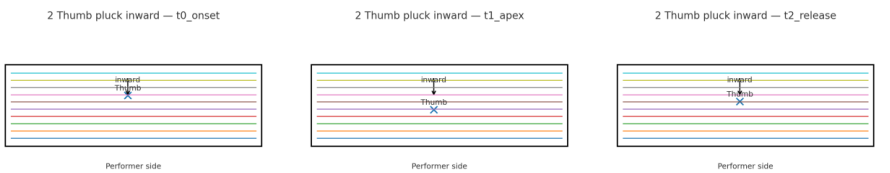

Fig. A.2. Thumb pluck inward (Tiao, thumb).

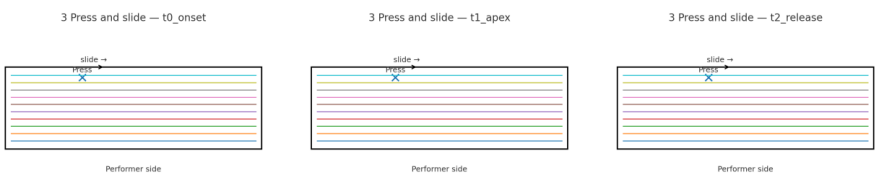

Fig. A.3. Press and slide (An-Hua).

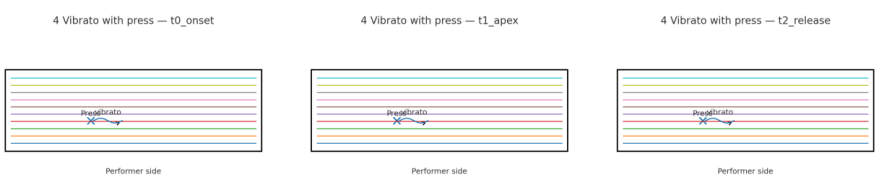

Fig. A.4. Vibrato with press (Yao-yin).

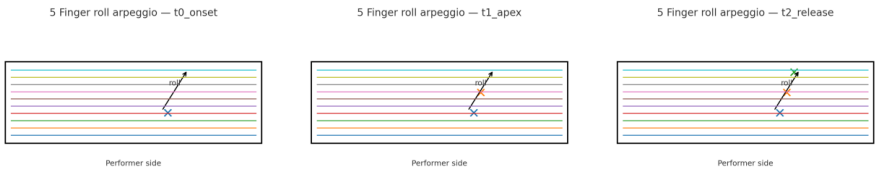

Fig. A.5. Finger roll arpeggio (Gun-zou).

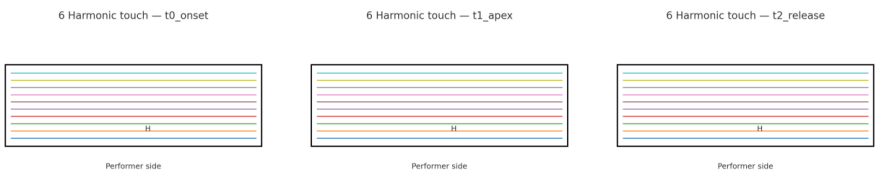

Fig. A.6. Harmonic touch (Fan-yin).

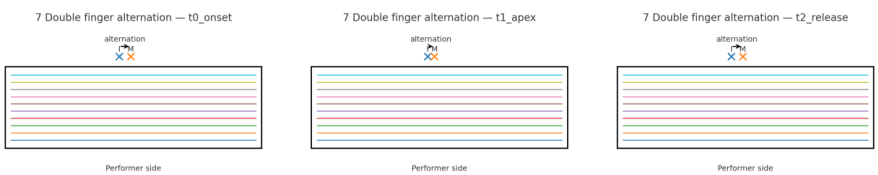

Fig. A.7. Double-finger alternation (Shuang-zhi alternation).

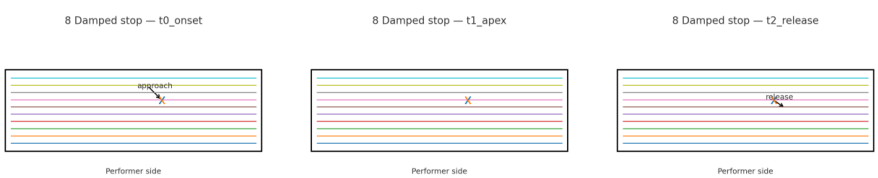

Fig. A.8. Damped stop (Mute/Stop).