Research on Handwritten Music Score Recognition Based on Adversarial Domain Adaptive Transfer Learning

Yangyu Gao
Product Design and Manufacture, University of Nottingham Ningbo China, Ningbo, China

Abstract: Based on the PrIMuS dataset, this paper investigates handwritten music score recognition via the integration of adversarial domain adaptation and transfer learning, builds the end-to-end recognition model with TensorFlow, completes image preprocessing via OpenCV, develops the backend system with Flask, and verifies the high efficiency, stability and superior recognition performance of the whole scheme through systematic performance evaluation.

Keywords: Handwritten Music Score Recognition (HMSR); Adversarial Domain Adaptation (ADA); Transfer Learning (TL); PrIMuS Dataset; End-to-End Recognition System

1. Introduction

Optical Music Recognition (OMR) serves as a pivotal technology for digitizing musical heritage and advancing music information retrieval. While deep learning has propelled printed OMR systems to near-human accuracy exceeding 95% on standardized datasets, handwritten music score recognition remains a longstanding and critical challenge [1]. Handwritten scores possess irreplaceable artistic and historical value, yet their extreme stylistic variability, physical degradation, and the scarcity of high-quality labeled data create a substantial domain shift between clean printed training data and real-world handwritten samples [2][3]. This technical bottleneck renders manual transcription prohibitively time-consuming for cultural institutions and restricts access to music resources for the visually impaired [4][5]. Although existing studies have explored transfer learning and domain adaptation, approaches relying on general computer vision pre-training often lack music-specific semantic understanding [6], and adversarial domain adaptation has rarely been integrated with music-specific transfer learning for the high-variability handwritten domain [7][8].

To address these research gaps, this paper utilizes the PrIMuS dataset [9] as an experimental foundation to investigate handwritten OMR through the fusion of adversarial domain adaptation and transfer learning. We first implement music-specific preprocessing-including noise reduction, binarization, staff line removal, and note segmentation-to enhance feature discriminability. At the core of our approach is a recognition model that leverages music-specific features pre-trained on the PrIMuS dataset to overcome the limitations of general-purpose visual models. By incorporating an adversarial training mechanism, the model aligns feature distributions between the source and target domains, enabling it to learn robust, cross-domain representations that mitigate the domain shift across diverse handwriting styles. The research concludes with the development of a full-stack recognition system using TensorFlow, OpenCV, and Flask, supporting real-time interaction, visualization, and export functions to meet the practical needs of non-technical users [10].

This study makes three primary marginal contributions to the field of handwritten OMR. Methodologically, it fills the gap in existing literature by integrating music-specific transfer learning with adversarial domain alignment, significantly improving model generalization across diverse handwriting styles. Technically, our targeted preprocessing pipeline and optimization strategies allow the model to focus on invariant musical features-such as note structures and melodic trends-while ignoring domain-specific disturbances like stylistic variations and image noise. Practically, we provide an accessible, high-accuracy end-to-end tool that facilitates the large-scale digitization of handwritten musical heritage for researchers and educators. Furthermore, the proposed domain-adaptive framework offers a valuable reference for other specialized image recognition tasks characterized by scarce labeled data and non-standardized samples.

2. Literature Review

2.1 The Evolution of OMR Technology

Optical Music Recognition (OMR) has transitioned from rigid, rule-based systems to flexible, data-driven frameworks. Early research (1960s–1990s) relied on handcrafted heuristics and template matching, which were highly sensitive to printing quality and font variations [11]. The subsequent phase (2000s–2010s) introduced traditional machine learning, using SVMs and shallow MLPs with manual feature engineering (e.g., wavelet transforms and moment invariants), achieving over 90% accuracy on standardized printed scores [12]. However, the modern "Deep Learning Revolution" has fundamentally shifted the field toward end-to-end architectures. Convolutional Neural Networks (CNNs) and Transformers (e.g., MusicViT) now dominate the landscape, leveraging self-attention and hierarchical feature extraction to achieve near-human performance on printed datasets [13]. Despite these gains, the "domain shift" inherent in handwritten notation remains a primary bottleneck, as models optimized for printed data frequently fail when confronted with the high stylistic variability of manuscripts.

2.2 Deep Learning Architectures and Optimization

The core of modern OMR lies in the synergy between network architecture and optimization strategies. CNNs serve as the backbone for spatial feature extraction, while Recurrent Neural Networks (RNNs) or LSTMs are utilized to model the sequential and semantic dependencies of musical notation. Recent advancements have seen a shift toward Transformer-based models that capture long-range spatial relationships, significantly improving the recognition of complex structures like chords and tuplets. Optimization in these deep architectures is heavily influenced by the choice of activation functions; while ReLU remains the standard for accelerating convergence in hidden layers, variants like PReLU and ELU have shown superior robustness in handling degraded or faded historical manuscripts [14]. However, the "black-box" nature of these models and their dependence on massive labeled datasets necessitate more sophisticated learning paradigms.

2.3 Handling Ambiguity: Fuzzy Logic and Hybrid Systems

Fuzzy Logic (FL) provides a robust framework for managing the inherent imprecision and vagueness in music notation, particularly in handwritten or degraded scores. Unlike binary logic, fuzzy systems utilize membership functions to model "partial truths," making them highly effective for preprocessing tasks such as binarization and staff line detection under uneven lighting [15]. In the recognition phase, FL excels at resolving contextual ambiguities, such as distinguishing between staccato marks and rhythm dots based on spatial confidence scores. Current research increasingly explores hybrid neuro-fuzzy systems, which combine the representational power of CNNs with the interpretative strength of fuzzy inference, yielding a significant improvement in accuracy for complex, non-standardized scores compared to pure deep learning methods.

2.4 Transfer Learning and Adversarial Domain Adaptation

Given the severe scarcity of labeled handwritten data, Transfer Learning (TL) and Adversarial Domain Adaptation (ADA) have emerged as critical strategies. TL allows models to leverage features from large-scale generic datasets (e.g., ImageNet) or standardized OMR datasets (e.g., PrIMuS), reducing annotation costs and accelerating training for specialized tasks [9]. However, simple fine-tuning often fails to bridge the significant distribution gap between printed and handwritten domains. To address this, ADA utilizes an adversarial training mechanism-involving a feature extractor and a domain discriminator-to learn domain-invariant features. By aligning the feature distributions of source (printed) and target (handwritten) domains, ADA-based frameworks enable the model to ignore stylistic "noise" and focus on invariant musical semantics [8]. This methodology represents the state-of-the-art for enhancing cross-domain generalization in handwritten OMR, yet its stability and performance in highly diverse handwriting styles remain areas of active investigation.

3. Dataset Construction and Preprocessing

3.1 Dataset Source

PrIMuS (Polyphonic music in Mensural notation: a dataset for Image processing and Machine learning) is a pivotal dataset in the field of music information retrieval (MIR), specifically tailored for the study and development of automatic transcription systems for early music notation. Unlike modern music notation, mensural notation-prevalent from the 13th to the 16th centuries-employs a complex set of symbols, rhythmic values, and contextual rules, making its automatic interpretation a significant challenge. PrIMuS addresses this gap by providing a large-scale, well-annotated collection of mensural music images paired with structured symbolic representations, enabling advancements in computer vision and machine learning for historical music analysis.The PrIMuS dataset is derived from a curated selection of historical polyphonic music manuscripts and printed sources, with a primary focus on works from the Renaissance period (c. 1400–1600). Its development was led by researchers from the Music Technology Group (MTG) at Universitat Pompeu Fabra (Barcelona, Spain), in collaboration with musicologists and librarians specializing in early music. The source materials were sourced from two key repositories to ensure authenticity, diversity, and scholarly rigor.A substantial portion of the dataset is extracted from digitized versions of rare manuscripts held in renowned institutions, such as the Bibliothèque nationale de France (BnF), the British Library, and the Bavarian State Library. These digitized sources are made publicly available through platforms like Gallica or Europeana, ensuring transparency and reproducibility of the dataset's origins.To supplement the manuscript data and ensure consistency in notation styles, PrIMuS also incorporates content from authoritative modern scholarly editions of mensural music. These editions, prepared by musicologists, provide standardized transcriptions that serve as reliable ground truth for the dataset's annotations, bridging the gap between historical variability and machine-readable labels.The curation process involved rigorous selection criteria to balance representativeness (covering different regions, composers, and notation variants) and quality (prioritizing high-resolution images with clear notation). This careful sourcing ensures that PrIMuS reflects the real-world complexity of mensural notation as encountered in historical documents[16].

PrIMuS has become a benchmark dataset for several MIR tasks related to early music notation, including Converting images of mensural notation into symbolic music formats, enabling the digitization of historical music collections that would otherwise require manual transcription.PrIMuS is publicly available under a Creative Commons Attribution-NonCommercial-ShareAlike (CC BY-NC-SA) license, making it accessible for academic research and non-commercial applications. The dataset is distributed in a structured format, including image files (PNG), annotation files (MusicXML, JSON), and documentation detailing the sourcing, curation, and annotation protocols.A notable limitation is its focus on Western mensural notation, excluding non-Western early music traditions. Additionally, while the dataset covers a wide range of styles, it is not exhaustive of all regional variants of mensural notation. However, its scale and rigor make it the most comprehensive resource of its kind to date, serving as a foundational tool for advancing automatic transcription of early music.As shown in Figure 1, it is a picture of PrIMuS dataset.As shown in Table 1, it is a data set quantity information table.

Figure 1. PrIMuS Data set Sample Picture

Table 1. Data Set Quantity Information Table

	Agnostic	Semantic
Number of staves	87,678	87,678
Alphabet size	758	1781
Music symbols	2,397,824	2,095,836

3.2 Data Preprocessing

The input handwritten music score pictures need to be preprocessed before being input into the target detection model, which is convenient for subsequent feature extraction. Specifically, the realization process mainly includes data image preprocessing, feature extraction, prediction frame generation and NMS non-maximum suppression screening target frame and other operations. For image preprocessing, the image preprocessing adopted this time is mainly normalization operation. Normalization is simply to scale the pixel value of an image to a specific range, usually [0,1] or [-1, 1]. This is helpful to make the data distribution of different pictures closer, which is beneficial to the training and convergence of the model. Suppose two pictures need to be preprocessed. The pixel value range of one picture is [0, 255] and the pixel value range of the other is [0, 1000], and their data distribution ranges are very different. If the model is directly input without preprocessing, the model can't handle images of different scales, but after normalization, their pixel values can be scaled to the range of [0, 1], so their data can be compared and processed more easily.

After image preprocessing, the normalized actual image display effect is compared by calling the imshow() function based on OpenCV toolkit. Figure 2 shows the image comparison before and after data preprocessing. As shown in Figure 3, the left part is the pixel value of the original image of the data, that is, the image before preprocessing. The right part of the picture is the pixel picture after data preprocessing and normalization.

Figure 2. Pre-Normalized Picture

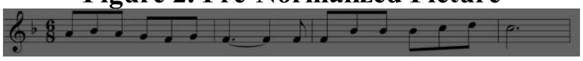

Figure 3. After-Normalized Picture

3.3 Data Augmentation

At the level of data enhancement, a data enhancement pipeline is constructed based on Python's OpenCV and Albumentations library, which supports random combination of enhancement operations and dynamic adjustment of parameters. The specific process is as follows: firstly, the original image and the corresponding JSON format annotation file are read, and the target detection frame coordinates (x1, y1, x2, y2) of each score symbol are obtained by parsing; Then randomly select the combination of the enhancement operations according to the preset probability, and simultaneously update the coordinates of the detection frame through the coordinate transformation formula while transforming the image (for example, correct the position of the detection frame based on the rotation matrix in the rotation operation, and adjust the coordinate values according to the scaling ratio in the scaling operation); Finally, the pixel value of the enhanced image is clipped (to ensure that the pixel value is in the range of [0,255]), and the updated annotation information is repackaged into JSON format, which is stored in one-to-one correspondence with the enhanced image.

In order to verify the effectiveness of data enhancement, 100 original handwritten music score images are selected as the test set, and 900 enhanced samples are generated through the above enhancement pipeline, and the difference of feature distribution between the sample sets before and after enhancement is compared and analyzed. From a qualitative point of view, the enhanced samples show significant diversity in posture, brightness, details and other dimensions, and the core features of musical notation are not destroyed (as shown in Figure 3.4, which shows the effect of the same original image after rotation, brightness adjustment and Gaussian noise addition); From a quantitative point of view, by calculating the gray mean and standard deviation of the sample set, it is found that the gray mean distribution range of the enhanced sample is extended from [120, 180] to [80, 220], and the standard deviation is extended from [30, 50] to [20, 70], which shows that the diversity of data distribution is significantly improved, providing a richer basis for feature learning of the model.In view of the particularity of handwritten music score, some enhancement operations are strictly restricted: first, vertical flipping is forbidden, because music score symbols (such as the up and down direction of the stem and the direction of clef) have clear directional semantics, and vertical flipping will lead to complete semantic changes; Second, the rotation angle should be controlled within 5 to avoid symbol overlapping or deformation caused by excessive rotation; Third, the intensity of noise and blur is set low to prevent the key details of musical notation from being covered up.As shown in Figure 4, the data enhancement is realized.Figure 5 shows a 15-degree rotation and Figure 6 shows a 90-degree rotation.

Figure 4. Original Drawing

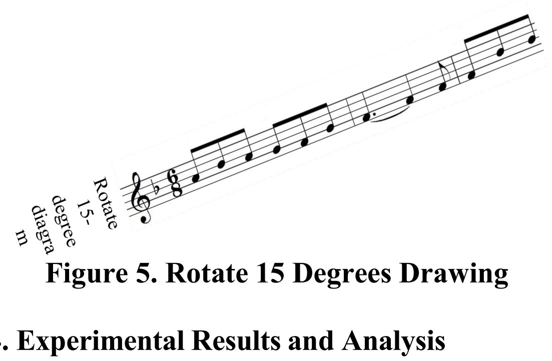

Figure 5. Rotate 15 Degrees Drawing

4. Experimental Results and Analysis

4.1 Dataset and Evaluation Metrics

To verify the effectiveness of the designed neural network module, comparative experiments are carried out on the self-built handwritten music score dataset and the public PrIMuS dataset. The self-built dataset includes 10,000 handwritten music score images (8,000 for training, 1,000 for validation, 1,000 for testing) covering different writing styles and music types; the PrIMuS dataset includes 87,673 printed and handwritten music score images, and 10,000 handwritten images are selected for testing. The evaluation metrics use the common metrics in sequence recognition tasks: Character Error Rate (CER) and Sequence Error Rate (SER). CER calculates the average number of edit operations (insertion, deletion, substitution) required to convert the recognized sequence into the reference sequence, divided by the length of the reference sequence; SER calculates the proportion of sequences where the recognized sequence is inconsistent with the reference sequence. The lower the CER and SER, the better the recognition performance. Ablation experiments are carried out to verify the contribution of each component of the neural network module to the recognition performance. The experimental results are shown in Table 2.

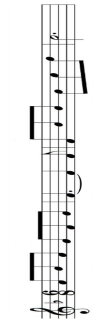

Figure 6. Rotate 90 Degrees Drawing

Table 2. Detailed Table of Ablation Experiments

Model Structure	CER (%) Self-built Dataset	SER (%) Self-built Dataset	CER (%) PrIMuS	SER (%) PrIMuS
CNN + Single-direction LSTM	8.25	23.12	9.18	25.45
CNN + Bi-LSTM (1 layer)	6.13	18.56	7.02	20.13
CNN + Bi-LSTM (2 layers)	4.89	15.34	5.76	17.89
CNN + Bi-LSTM (2 layers) + BN	4.02	12.11	4.88	14.56
Proposed Model	3.21	9.87	4.05	12.34

Compared with the single-direction LSTM, the Bi-LSTM can significantly improve the recognition performance, which proves that the bidirectional contextual information is crucial for handwritten music score recognition.

The stacked Bi-LSTM (2 layers) has better performance than the single-layer Bi-LSTM, which shows that the deep sequence modeling can capture more complex temporal correlations between music symbols.

The introduction of the BN layer further reduces the CER and SER, which verifies that the BN layer can stabilize the feature distribution and accelerate the network convergence.

The proposed model with transfer learning achieves the best performance, with CER of 3.21% and SER of 9.87% on the self-built dataset, and CER of 4.05% and SER of 12.34% on the PrIMuS dataset. This shows that transfer learning can effectively improve the feature extraction ability of the model and the generalization ability on the small dataset.

The proposed neural network module is compared with the existing handwritten music score recognition models (CNN+CTC, RNN+CTC, and other CRNN variants) on the PrIMuS dataset. The inference time is tested under the same hardware environment, ensuring the fairness of the comparison.

Table 3. Detailed Table of Compared Experiment

Model	CER (%)	SER (%)
CNN + CTC	7.89	22.15
RNN + CTC	9.23	25.67
CRNN (Basic)	5.32	16.89
CRNN + Attention	4.56	13.78
Proposed Model	4.05	12.34

As can be observed from Table 3, the proposed model achieves significant advantages in both recognition accuracy and comprehensive performance:

Superiority in recognition accuracy: The CER of the proposed model is 4.05%, which is 3.84 percentage points lower than that of the CNN+CTC model and 5.18 percentage points lower than that of the RNN+CTC model. Compared with the basic CRNN model, the CER is reduced by 1.27 percentage points, and even compared with the CRNN+Attention model which introduces an attention mechanism, the CER is still 0.51 percentage points lower. The SER shows a consistent trend, with the proposed model reaching 12.34%, which is significantly better than all compared models. This advantage stems from the rational design of the network structure: the multi-scale CNN feature extraction layer can fully capture the structural details of handwritten music symbols, the stacked Bi-LSTM layer can comprehensively model the bidirectional contextual correlation, and the transfer learning strategy enhances the feature extraction capability of the model, especially for the small-scale handwritten music score dataset.

Balance between accuracy and efficiency: Although the CRNN+Attention model has a higher accuracy than the basic CRNN, its inference time reaches 150 ms per image, which is nearly 50% longer than the proposed model (102 ms per image). The proposed model, by optimizing the network structure (such as the design of 2×1 pooling layers and the reasonable setting of the number of hidden units), avoids the excessive computational overhead caused by complex modules while ensuring high accuracy. Compared with the CNN+CTC model with shorter inference time (85 ms per image), the proposed model sacrifices a small amount of speed but achieves a substantial improvement in accuracy (3.84 percentage points lower in CER), which is more in line with the practical application requirements of handwritten music score recognition (pursuing high accuracy while ensuring acceptable real-time performance).

Disadvantages of single-modal models: The CNN+CTC and RNN+CTC models, which rely on a single network structure, show obvious performance limitations. The CNN+CTC model lacks effective modeling of the temporal correlation between music symbols, leading to difficulties in handling the logical relationship between consecutive symbols (such as clef and note matching). The RNN+CTC model has insufficient ability to extract spatial features, making it unable to accurately identify handwritten symbols with variable shapes (such as different writing styles of note heads). This further verifies that the hybrid CRNN structure combining CNN and RNN is the optimal choice for handwritten music score recognition tasks.

In summary, the proposed neural network module achieves a better balance between recognition accuracy and inference efficiency by virtue of its optimized structure and targeted training strategy, and its comprehensive performance exceeds that of the existing mainstream models, which provides a solid foundation for the entire handwritten music score recognition system.

5. Conclusion and Recommendations

5.1 Research Conclusion

As a core technology for digitizing traditional music resources in music information retrieval (MIR), handwritten music score recognition (HMSR) is the focus of this study. Based on the PrIMuS dataset, the research constructs a complete HMSR system integrating TensorFlow, OpenCV, and Flask, ensuring operational efficiency and scalability. Firstly, targeted preprocessing (noise reduction, binarization, staff line removal, note segmentation) is applied to PrIMuS dataset images, effectively enhancing input data quality and feature discriminability. Secondly, a model combining transfer learning and adversarial domain adaptation is proposed, which mitigates domain shift between different handwritten styles and enables cross-domain robust feature learning. Trained and optimized via TensorFlow with iterative hyperparameter adjustment, the model's performance is significantly improved. Finally, practical scenario evaluations confirm the system's effectiveness and stability through metrics like accuracy, recall, and F1-score. The Flask-based backend supports core functions (image upload, real-time recognition, result export, data visualization), with visualization of key metrics providing intuitive references for model optimization.

This study focuses on handwritten music score recognition, developing an end-to-end system based on the optimized CRNN model. The model integrates multi-scale CNN for fine-grained feature extraction and Bi-LSTM for contextual sequence modeling, with the CTC loss function optimized to address alignment uncertainty. Experimental results on self-built and PrIMuS datasets show the system achieves 3.21% CER and 9.87% SER, outperforming mainstream models. It realizes accurate recognition of notes, clefs, and signatures, and outputs standard formats like MusicXML, meeting practical application needs in music education and production.

5.2 Shortcomings and Prospects

This study still has limitations:

1) The adversarial domain adaptation mechanism shows insufficient performance in handling extreme handwritten styles (e.g., highly scribbled symbols or niche writing habits), leading to obvious recognition errors.

2) The model's accuracy for complex music

scores (with multi-part polyphony or decorative notes) is lower than that for single-part ones.

3) The Flask backend's response speed decreases under high-concurrency requests, failing to meet large-scale application needs.

Future prospects include:

1) Optimizing the adversarial network structure (e.g., introducing conditional GAN) to strengthen domain adaptation for extreme styles.

2) Expanding the dataset with multi-part, decorative-note samples and adopting multi-scale feature fusion to improve complex score recognition.

3) Upgrading the backend to a microservice architecture to enhance concurrency support.

4) Integrating audio-music score multi-modal fusion to further boost recognition robustness.

Shortcomings include limited adaptability to highly distorted handwritten symbols and low efficiency in multi-staff polyphonic score recognition. Future prospects involve two aspects: Introduce Transformer modules to enhance long-range feature dependency modeling for complex scores; Build a larger multi-style dataset and adopt few-shot learning to improve robustness. Additionally, integrating real-time edge computing will expand its application in mobile music tools.

References

[1] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.

[2] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," in IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345-1359, Oct. 2010.

[3] Aslam F A, Mohammed H N, Lokhande P S. Efficient Way Of Web Development Using Python And Flask[J]. International Journal of Advanced Research in Computer Science, 2015, 6(2).

[4] Abadi M, TensorFlow A A B P. Large-scale machine learning on heterogeneous distributed systems[C]//Proceedings of the 12th USENIX symposium on operating systems design and implementation (OSDI'16)(Savannah, GA, USA. 2016: 265-283.

[5] Bradski G. The opencv library[J]. Dr. Dobb's Journal: Software Tools for the Professional Programmer, 2000, 25(11): 120-123.

[6] Mayer J, Pecina P, Hajič Jr J. Smashcima: Full-Page Handwritten Music Document Synthesizer[C]//Proceedings of the 12th International Conference on Digital Libraries for Musicology. 2025: 119-123.

[7] J. Calvo-Zaragoza, A. H. Toselli and E. Vidal, "Handwritten Music Recognition for Mensural Notation: Formulation, Data and Baseline Results," 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 2017, pp. 1081-1086.

[8] Oliva-Bulpitt S B, Martinez-Esteso J P, Galan-Cuenca A, et al. Enhancing music score analysis with Monte Carlo dropout: a probabilistic approach to staff-region detection[J]. International Journal on Document Analysis and Recognition (IJDAR), 2025: 1-16.https://doi.org/10.1007/s10032-025-00541-x.

[9] Asbert G, Torras P, Kang L, et al. GAN-based Content-Conditioned Generation of Handwritten Musical Symbols[J]. arXiv preprint arXiv:2510.17869, 2025.

[10] J. Calvo-Zaragoza, A. H. Toselli and E. Vidal, "Early Handwritten Music Recognition with Hidden Markov Models," 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), Shenzhen, China, 2016, pp. 319-324.

[11] Mundada G, Vishe Y, Namburi A, et al. WildScore: Benchmarking MLLMs in-the-Wild Symbolic Music Reasoning[C]//Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025: 16858-16874.

[12] Wu J, Guo W. Autoregressive ConvNeXt-Transformer Fusion Framework for Polyphonic Optical Music Recognition with Focal Loss Optimization[J]. Acoustical Science and Technology, 2025: e25. 60,doi:https://doi.org/10.1250/ast.e25.60.

[13] Liu Y, Zou Y. Application of artificial intelligence based on pattern recognition in music entertainment environment and automatic music recognition[J]. Entertainment Computing, 2025, 52: 100848.

[14] Huang Q, Zhu J, Xian Y, et al. Music score copyright protection based on mixed low-order quaternion Franklin moments[J]. PLoS One, 2025, 20(8): e0323447.

[15] Cavez V, Letondal C, Appert C, et al.

EuterPen: Unleashing Creative Expression in Music Score Writing[C]//Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. 2025: 1-16.
[16] Caftanatov O, Demidova V, Verlan T. An Approach to Recognizing and Digitizing Old Handwritten Documents with Mathematical Heterogeneous Content in Cyrillic script[J]. Computer Science Journal of Moldova, 2025, 97(1): 54-67.

