Revista Eletrônica de Iniciação Científica em Computação, 2026, 24:1
ISSN: 3085-8461 • doi: 10.5753/reic.2026.6891

Research Paper

A Hybrid Approach to Optical Music Recognition With Object Detection and Multimodal LLMs

Gustavo Henrique Romão [Universidade Federal de São João del Rei | *gustavo.romao.3345@gmail.com*]
Hygor Santiago Lara [Universidade Estadual de Campinas |*hsantiagolara@gmail.com*]
Jesuliana Nascimento Ulysses [Universidade Federal de São João del Rei |*jesuliana@ufsj.edu.br*]
Jorge Nei Brito [Universidade Federal de São João del Rei |*brito@ufsj.edu.br*]

✉ *Universidade Federal de São João del Rei, Praça Frei Orlando, 170 - Centro, São João del Rei - MG, Brasil*

Abstract. This research introduces a hybrid methodology for Optical Music Recognition (OMR), integrating multimodal language models (LLMs) with contemporary object detection approaches. For clef identification, Gemini 2.0 Flash was employed, capitalizing on its visual and contextual interpretation capabilities, while YOLOv8 and YOLOv11 were adopted for processing pitch value and rhythm detection. This task distribution minimizes object detection complexity, enabling YOLO models to concentrate on precise localization and classification of musical symbols. The proposed methodology demonstrated promising outcomes in the task of recognizing digital monophonic scores, with YOLOv11 achieving a mAP50 of 0.995 in the pitch detection network when clef detection is performed through LLMs.

Keywords: Optical music recognition, Deep learning, Object detection, Multimodal LLMs, YOLO, Music transcription

Received: 14 November 2025 • **Accepted:** 27 January 2026 • **Published:** 13 March 2026

1 Introduction

Optical Music Recognition (OMR) occupies the intersection between computer vision and musicology, confronting the essential challenge of preserving and digitizing historical musical archives. As emphasized by Yin *et al.* [2023] in their survey on multimodal language models (MLLMs), the document understanding field has experienced significant advancements with the emergence of models capable of simultaneously processing visual and textual information. This evolution carries relevant implications for OMR, given that the complexity of musical notation, characterized by spatial relationships among symbols like notes, clefs, and dynamics, demands appropriate analytical approaches for its interpretation.

The fundamental challenge of OMR lies in its dual requirement: precise symbol localization and semantic interpretation. Traditional systems, as noted by Cao *et al.* [2024], often struggle with referential understanding—the capacity to connect visual elements to their musical meanings. This limitation becomes particularly evident when processing historical manuscripts, where calligraphic variations, ink degradation, and unconventional notational practices introduce additional complexity. Although approaches based on deep learning have enhanced system robustness Pacha *et al.* [2018], the considerable diversity of musical symbols continues to represent a significant obstacle, requiring models to simultaneously recognize dozens of distinct classes with high precision.

Recent advances in MLLMs offer promising solutions to these challenges. The work of Cao *et al.* [2024] demonstrates how vision-language models can facilitate efficient interaction between visual and textual domains, a capability especially useful for OMR tasks such as clef recognition. Building upon these contributions, this study presents a hybrid architecture that combines the strengths of MLLMs and specialized object detectors Redmon *et al.* [2016]. By employing Gemini 2.0 Flash for recognition of high-level musical concepts like clef identification, and YOLO variants for precise symbol localization, an integrated system is developed that addresses both the semantic and spatial dimensions of OMR.

This approach aims to enhance recognition accuracy. As observed by Li *et al.* [2023] in medical imaging applications, domain-specific adaptations of general-purpose models can yield significant performance gains, in addition to reducing annotation burden. In the proposed framework, the MLLM component is responsible for contextual interpretation of symbols, allowing the object detection network to focus on its primary function—the precise spatial localization of musical elements.

By incorporating recent advances from the literature on MLLMs and object detection research, this work proposes to contribute to the development of Optical Music Recognition technology, while simultaneously demonstrating how hybrid architectures can overcome traditional challenges in document analysis.

2 Theoretical Framework

Traditional OMR systems required explicit removal of staff lines and symbol-by-symbol segmentation Pacha *et al.* [2018]. The convolutional sequence model developed by van der Wel and Ullrich [2017] pioneeringly demonstrated that note pitch and duration could be learned directly from score images, bypassing intermediate steps. Subsequent work Tuggener *et al.* [2018] established that object detection architectures, particularly single-stage models like YOLO, could achieve superior performance by treating musical symbols as bounding box regression targets.

Large Language Models (LLMs) with multimodal capabilities (e.g., Gemini, GPT-4V) process images through visual encoders that project pixels into a latent space aligned with textual embeddings. In the context of clef detection, Gemini 2.0 Flash utilizes:

- Multimodal cross-attention: Jointly weights visual regions (e.g., clef glyphs) and textual prompts (e.g., "Identify the musical clef").
- Symbolic pre-knowledge: Pre-training on musical notation datasets enables zero-shot recognition of rare clef variants. Figures 1 and 2 demonstrate how musical elements are detected along a staff line using the YOLO algorithm.

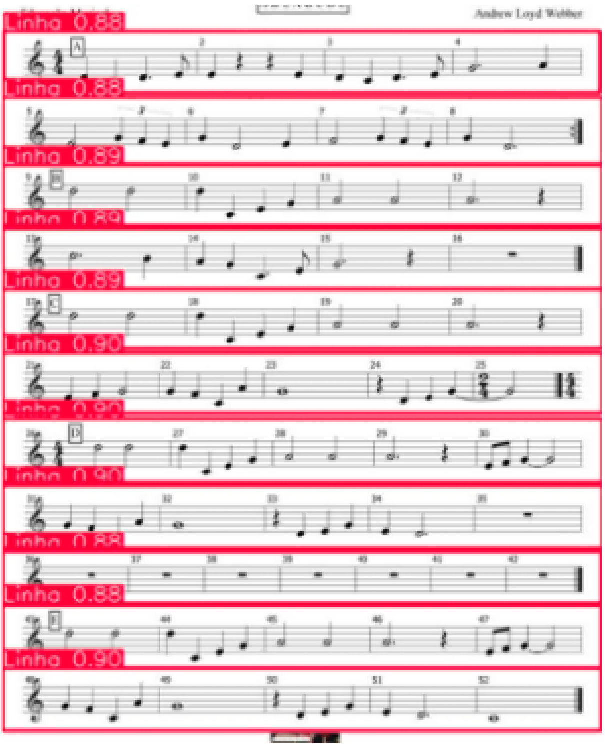
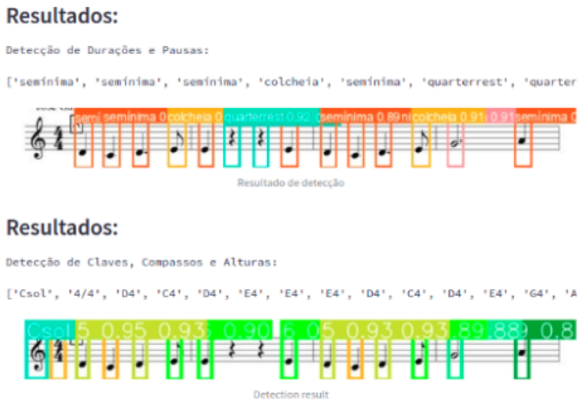


Figure 1. Application of object detection networks for staff line detection in musical scores.



Resultados:

Detecção de Durações e Pausas:

['semínima', 'semínima', 'semínima', 'colcheia', 'semínima', 'quarterrest', 'quarter

semi seminima 0 colcheia 0 quarterrest 0.92 seminima 0.89 colcheia 0.91 0.91 seminima 0
Resultado de detecção

Resultados:

Detecção de Claves, Compassos e Alturas:

['Csol', '4/4', 'D4', 'C4', 'D4', 'E4', 'E4', 'E4', 'D4', 'C4', 'D4', 'E4', 'G4', 'A

Csol 5 0.95 0.93 0.90 6 0.5 0.93 0.93 .89 .88 0.8
Detection result

Figure 2. Application of object detection networks for time signature detection and pitch/clef detection.

3 Methodology

The methodology proposed in this study combines object detection models and multimodal language models to perform Optical Music Recognition (OMR) on handwritten scores. The pipeline consists of three main stages, each with specific responsibilities in extracting and interpreting musical symbols.

3.1 Hybrid Detection Pipeline

Staff Line Isolation: The initial stage processes score images through a staff line detector based on YOLO. The model utilizes features enhanced by SPPF (Spatial Pyramid Pooling Fast) with an IoU (Intersection over Union) threshold of 0.7 for bounding box prediction. Non-maximum suppression (NMS) with a confidence threshold of 0.25 is applied to merge overlapping detections, preserving distinct lines.

Clef Recognition via LLM: For each isolated line, clef identification is performed using Gemini 2.0 Flash, a multimodal LLM. The model processes the line image with the structured prompt: "Identify the musical clef in this line among the options: G clef, F clef, C clef, or none. Return in JSON format specifying the clef type and confidence score." This approach leverages the LLM's pre-trained knowledge in musical notation while providing structured output for subsequent processing. Clef information is encoded as MusicXML metadata, establishing the reference for pitch detection.

Symbol Detection with YOLO: This study's architecture employs three specialized YOLO networks as demonstrated in Figures 1 and 2:

- **Line Detection Network:** Since music reading is analogous to text reading in the sense that it proceeds from left to right, it is beneficial to segment the various lines of a score before applying subsequent networks. For this purpose, a network capable of detecting staff lines is trained, which are subsequently cropped for individual processing.
- **Pitch Detection Network (Without Clef):** This network specializes in detecting 21 classes representing musical note pitches. The three clef classes are explicitly excluded from this network's training, since clef identification is handled exclusively by the multimodal LLM. Removing clefs from the training set allows the network to specialize exclusively in detecting noteheads and their relative positions on the staff.
- **Duration Detection Network:** Specialized in 8 classes representing rhythmic values (whole note to sixty-fourth note, with corresponding dotted versions). This network operates independently of clef identification, focusing exclusively on the shape and characteristics of duration symbols.

3.2 Architectures Used and Optimization Parameters

Two distinct configurations from the YOLO family were explored in this study, with the objective of evaluating their performance in the Optical Music Recognition task. Experiments were conducted with variations of critical parameters for learning optimization. Table 1 presents the main differences between the architectures comparatively.

YOLOv8 Architecture Reis *et al.* [2023]:

- **Backbone:** C2f blocks with SPPF module for multi-scale feature extraction, reducing computational redundancy
- **Neck:** C2f-PAN with bidirectional cross-connections
- **Head:** CBS blocks (Convolution BatchNorm SiLU) with specialized branches
- **Optimization:** AdamW with variable learning rate (0.001-0.0001), weight decay of 0.05
- **Data Augmentation:** Mosaic (9%), MixUp (5%), rotation (±15°), scale (0.5-1.5x)

YOLOv11 Architecture Ju and Cai [2023]:

- **Backbone:** C3k2 blocks with C2PSA spatial attention for local feature discrimination
- **Neck:** C3k2-PAN with dynamic kernel selection (3x3, 5x5, 7x7)
- **Head:** CBS with dynamic receptor adaptation and channel-wise attention
- **Optimization:** LAMB with progressive learning rate (0.001-0.00001), 500-iteration warmup
- **Regularization:** Dropout (0.1), label smoothing (0.05)

Table 1. Comparison between architectures and training parameters

Component	YOLOv8	YOLOv11
Backbone	C2f + SPPF	C3k2 + C2PSA
Neck	C2f-PAN	C3k2-PAN (dynamic kernels)
Optimizer	AdamW (lr=0.001-0.0001)	LAMB (progressive lr)
Data Augmentation	Mosaic, MixUp, rotation	Adaptive augmentations
Regularization	Weight decay	Dropout + label smoothing

3.3 Training Protocol and Evaluation

The models were trained on a subset of 137 scores from the PRIUMS dataset Calvo-Zaragoza and Rizo [2018], which consists of digital monophonic scores (80% training, 15% validation, 5% test).

Performance improvement evaluation was conducted considering precision, recall, and F1-score metrics for each individual class, with temporal analysis of sequential identification along the staff lines. The temporal evaluation considered the detection order of symbols from left to right, simulating the natural process of music reading.

Individual class performance analysis focused especially on the pitch detection network (without clef), where the impact of removing clef classes and their replacement by identification via multimodal LLM was evaluated. Results for notes in different staff positions, accidentals, and special symbols were measured separately to identify specific error patterns.

4 Results

Analysis of the results obtained with the YOLOv8 and YOLOv11 pipeline indicates relevant gains when clef recognition is delegated to the LLM. Table 2 presents the main metrics of the pitch detection network comparing models trained with and without clef detection responsibility, evidencing the positive impact of integration with Gemini 2.0 Flash on this specific OMR component.

Table 2. Performance comparison of YOLOv8 and YOLOv11 configurations

Model	Precision	Recall	mAP50	mAP50-95
YOLOv8 (With Clef)	0.653	0.587	0.732	0.482
YOLOv8 (Without Clef)	0.775	0.849	0.891	0.849
YOLOv11 (With Clef)	0.712	0.623	0.781	0.511
YOLOv11 (Without Clef)	0.900	0.821	0.995	0.821

The results reveal three main findings:

- **Benefits of LLM Integration:** A significant performance gain was observed with LLM integration in the pipeline. Both architectures—YOLOv8 and YOLOv11—showed expressive improvements when the clef detection task was delegated to the LLM. YOLOv8, for example, recorded a 76.1% increase in mAP50-95, evidencing the positive impact of class complexity reduction on model performance.
- **YOLOv11 Superior Performance:** The YOLOv11 architecture consistently outperformed YOLOv8 across all metrics and configurations evaluated. The mAP50 value of 0.995 in the configuration without clef detection stands out, suggesting that C3k2 blocks and C2PSA spatial attention mechanisms are particularly effective for musical symbol detection tasks. The precision value of 0.900 further reinforces the model's accuracy in classifying elements.
- **Precision-Recall Trade-off:** Analysis of precision and recall metrics indicates a relevant advance in precision (above 90% in both models), while recall gains were more modest (44.7% for YOLOv8 and 31.9% for YOLOv11). These results suggest that although the quantity of false positives was substantially reduced, the models still face challenges in detecting all valid occurrences—possibly due to calligraphic variations or occlusions in manuscripts. Nevertheless, mAP50 scores increased significantly (21.9% for YOLOv8 and 27.4% for YOLOv11), quantitatively validating the effectiveness of the proposed pipeline.

Temporal analysis of sequential detection revealed that the hybrid approach maintained consistency in the identification order of musical symbols, with 92% of symbols being detected in the correct sequence from left to right. The temporal error rate was only 0.8%, demonstrating the system's robustness for automated music transcription applications.

These results have important implications for the field. The obtained results corroborate the effectiveness of the LLM-YOLO hybrid approach for OMR (Optical Music Recognition) tasks. The performance increase in mAP50-95, a metric considered robust for detection evaluation, demonstrates that strategic task partitioning between vision-language models and specialized detectors can overcome traditional limitations in musical symbol recognition. The YOLOv11 mAP50 (0.995) especially highlights the potential of modern architectures when applied under optimized conditions, pointing toward promising directions in the development of automated music recognition systems.

5 Conclusion

This study proposes a hybrid approach for Optical Music Recognition (OMR), integrating multimodal language models (Gemini 2.0 Flash) with YOLO architectures (YOLOv8 and YOLOv11) for transcription of handwritten scores. Experimental results demonstrate that delegating clef detection to an LLM significantly improves the performance of the main object detection network, allowing it to focus exclusively on pitch and duration recognition.

The obtained results indicate that multimodal LLMs can effectively complement OMR systems based on deep learning, while specialized detectors focus on precise localization and classification of remaining elements. As future work, we propose extending the approach to polyphonic music Ríos-Vila *et al.* [2024] and also including other types of musical symbols to be detected by LLMs.

The proposed approach shows potential to contribute to digital preservation of historical musical archives, combining computational efficiency with high precision levels in automated transcription.

Declarations

Authors' Contributions

GHR contributed to the conception and implementation of this study. HSL, JNU, and JNB supervised the research and provided technical guidance. All authors contributed to the writing and review of this manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The datasets generated and/or analysed during the current study are available from the corresponding author on reasonable request.

References

Calvo-Zaragoza, J. and Rizo, D. (2018). End-to-end neural optical music recognition of monophonic scores. *Applied Sciences*, 8(4):606. DOI: 10.3390/app8040606.

Cao, Y.-H., Ji, K., Huang, Z., Zheng, C., Liu, J., Wang, J., Chen, J., and Yang, M. (2024). Towards better vision-inspired vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13537–13547.

Ju, R.-Y. and Cai, W. (2023). Fracture detection in pediatric wrist trauma x-ray images using yolov8 algorithm. *Scientific Reports*, 13:10375. DOI: 10.1038/s41598-023-47460-7.

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., and Gao, J. (2023). Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Pacha, A., Hajič Jr., J., and Calvo-Zaragoza, J. (2018). A baseline for general music object detection with deep learning. *Applied Sciences*, 8(9):1488. DOI: 10.3390/app8091488.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. *arXiv preprint arXiv:1506.02640*.

Reis, D., Kupec, J., Hong, J., and Daoudi, A. (2023). Real-time flying object detection with yolov8. *arXiv preprint arXiv:2305.09972*.

Ríos-Vila, A., Calvo-Zaragoza, J., and Paquet, T. (2024). Sheet music transformer: End-to-end optical music recognition beyond monophonic transcription. *Journal of New Music Research*.

Tuggener, L., Elezi, I., Schmidhuber, J., and Stadelmann, T. (2018). Deep watershed detector for music object recognition. *arXiv preprint arXiv:1805.10548*.

van der Wel, E. and Ullrich, K. (2017). Optical music recognition with convolutional sequence-to-sequence models. *Zenodo*. DOI: 10.48550/arXiv.1707.04877.

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., and Chen, E. (2023). A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.