

Abordagem Híbrida para Reconhecimento Ótico de Música com Detecção de Objetos e LLMs Multimodais

Gustavo Henrique Romão¹, Hygor Santiago Lara², Jesuliana Nascimento Ulysses¹

¹¹Universidade Federal de São João del Rei

Praça Frei Orlando, 170 - Centro, São João del Rei - MG, Brasil

gustavo.romao.3345@gmail.com, jesuliana@ufsj.edu.br

²Universidade Estadual de Campinas (Unicamp)

Rua Mendeleyev, 200, Campinas - SP, Brasil

hsantiagolara@gmail.com

Resumo. Este estudo propõe uma abordagem híbrida para OMR, integrando modelos multimodais de linguagem (LLMs) com técnicas modernas de detecção de objetos. Para o reconhecimento de claves foi utilizado o Gemini 2.0 Flash, aproveitando sua capacidade de interpretação visual e contextual, enquanto YOLOv8 e YOLOv11 foram adotados para o processamento da detecção de valores de altura e ritmo. Esta divisão de tarefas reduz a complexidade da detecção de objetos, permitindo que os modelos YOLO se concentrem na localização precisa e classificação de símbolos musicais. A metodologia proposta apresentou resultados satisfatório na tarefa de reconhecimento de partituras monofônicas digitais, com o YOLOv11 atingindo um mAP50 de 0.995 quando a detecção de claves é realizada por meio de LLMs. reconhecimento óptico de música, aprendizado profundo, detecção de objetos, LLMs multimodais, YOLO, transcrição musical

1. Introdução

O Reconhecimento Óptico de Música (OMR) posiciona-se na fronteira entre visão computacional e musicologia, enfrentando o desafio crucial de preservar e digitalizar acervos musicais históricos. Como destacado por [1] em seu levantamento sobre modelos multimodais de linguagem (MLLMs), a área de compreensão documental passou por avanços importantes com o surgimento de modelos capazes de processar informações visuais e textuais simultaneamente. Essa evolução tem implicações relevantes para o OMR, uma vez que a complexidade da notação musical - caracterizadas por relações espaciais entre símbolos como notas, claves e dinâmicas - exige abordagens analíticas adequadas para sua interpretação.

O desafio fundamental do OMR reside em sua dupla exigência: localização precisa de símbolos e interpretação semântica. Sistemas tradicionais, como observado por [2], frequentemente apresentam dificuldades com o entendimento referencial - a capacidade de conectar elementos visuais aos seus significados musicais. Essa limitação torna-se especialmente evidente ao processar manuscritos históricos, onde variações caligráficas, degradação da tinta e práticas não convencionais de notação introduzem complexidade adicional. Embora abordagens baseadas em aprendizado profundo tenham aumentado a robustez dos sistemas [3], a grande

diversidade de símbolos musicais continua representando um obstáculo significativo, exigindo que os modelos reconheçam simultaneamente dezenas de classes distintas com alta precisão.

Avanços recentes em MLLMs oferecem soluções promissoras para esses desafios. O trabalho de [2] demonstra como modelos visão-linguagem podem promover uma interação eficiente entre domínios visuais e textuais - capacidade especialmente útil para tarefas de OMR, como reconhecimento de claves. Baseando-se nessas contribuições, este estudo apresenta uma arquitetura híbrida que combina os pontos fortes de MLLMs e detectores de objetos especializados [4]. Ao empregar o Gemini 2.0 Flash para reconhecimento de conceitos musicais de alto nível, como identificação de claves, e variantes do YOLO para localização precisa de símbolos, desenvolve-se um sistema integrado que aborda tanto as dimensões semânticas quanto espaciais do OMR.

Essa abordagem visa melhorar a acurácia do reconhecimento. Conforme observado por [5] em aplicações de imageamento médico, adaptações específicas de domínio para modelos de propósito geral podem gerar ganhos significativos de desempenho, além de reduzirem a carga de anotação. Na estrutura proposta, o componente MLLM é responsável pela interpretação contextual de símbolos, permitindo que a rede de detecção de objetos se concentre em sua principal função - a localização espacial precisa de elementos musicais.

Ao incorporar avanços recentes da literatura sobre MLLMs e da pesquisa em detecção de objetos, este trabalho propõe contribuir para o desenvolvimento da tecnologia de Reconhecimento Ótico de Música, ao mesmo tempo em que demonstra como arquiteturas híbridas podem superar desafios tradicionais na análise documental.

2. Referencial Teórico

2.1. Evolução dos Pipelines de OMR

Sistemas tradicionais de OMR exigiam a remoção explícita das linhas da pauta e segmentação símbolo a símbolo [3]. O modelo para sequência convolucional desenvolvido por [6] demonstrou pioneiramente que altura e duração das notas poderiam ser aprendidas diretamente de imagens

de partituras, contornando etapas intermediárias. Trabalhos subsequentes [7] estabeleceram que arquiteturas de detecção de objetos - particularmente modelos de estágio único como YOLO - poderiam alcançar desempenho superior ao tratar símbolos musicais como alvos de regressão de caixas delimitadoras.

2.2. Modelos Multimodais de Linguagem no Reconhecimento Visual

Modelos de Linguagem de Grande Porte (LLMs) com capacidades multimodais (por exemplo: Gemini, GPT-4V) processam imagens através de codificadores visuais que projetam pixels em um espaço latente alinhado com embeddings textuais. No contexto da detecção de claves, o Gemini 2.0 Flash utiliza:

- Atenção cruzada multimodal: Pondera conjuntamente regiões visuais (ex.: glifos de claves) e prompts textuais ("Identifique a clave musical").
- Pré-conhecimento simbólico: Pré-treinamento em conjuntos de dados de notação musical permite reconhecimento zero-shot de variantes raras de claves.

Resultados:

Deteção de Durações e Pausas:

['seminima', 'seminima', 'seminima', 'colcheia', 'seminima', 'quarterrest', 'quarter



Resultado de detecção

Resultados:

Deteção de Claves, Compassos e Alturas:

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



Detection result

Figura 1: Aplicação da detecção das redes de detecção de objetos.

3. Metodologia

A metodologia proposta neste estudo combina modelos de detecção de objetos e modelos multimodais de linguagem para realizar o Reconhecimento Óptico de Música (OMR) em partituras manuscritas. O pipeline é composto por três etapas principais, cada uma com responsabilidades específicas na extração e interpretação dos símbolos musicais.

3.1. Pipeline Híbrido de Detecção

Isolamento de Linhas da Pauta: A etapa inicial processa imagens de partituras através de um detector de linhas da pauta baseado em YOLO. O modelo utiliza recursos aprimorados por SPPF (Spatial Pyramid Pooling Fast) com um limiar de IoU (Intersecção sobre União) de 0.7 para predição de bounding boxes. É aplicada supressão não máxima (NMS) com limiar de confiança de 0.25 para mesclar detecções sobrepostas, preservando linhas distintas.

Reconhecimento de Claves via LLM: Para cada linha isolada, a identificação de claves é realizada usando Gemini 2.0 Flash, um LLM multimodal. O modelo processa a imagem da linha com o prompt estruturado: "Identifique a clave musical nesta linha entre as opções: sol (clave de G), fá (clave de F), dó (clave de C) ou nenhuma. Retorne em formato JSON especificando o tipo de clave e score de confiança." Esta abordagem aproveita o conhecimento pré-treinado do LLM em notação musical enquanto fornece saída estruturada para processamento posterior. As informações de clave são codificadas como metadados MusicXML, estabelecendo a referência para detecção de alturas.

Detecção de Símbolos com YOLO: A arquitetura deste estudo faz uso de três redes YOLO especializadas conforme demonstrada na figura 2:

- Rede de Detecção de linhas: Como a leitura de música é feita de maneira análoga à de texto, no sentido de que é feita da esquerda a direita, é de interesse segmentar as várias linhas de uma partitura antes da aplicação das redes subsequentes descritas abaixo. Para isso, é treinada uma rede capaz de detectar as linhas da partitura, para que sejam posteriormente recortadas, e, por fim, terem futuras redes aplicadas sobre elas.
- Rede de Detecção de Alturas: 24 classes rotuladas (21 sendo representativas das alturas das partituras e as três claves) incluindo cabeças de nota. As três classes de claves são explicitamente excluídas, sendo tratadas por LLM no treinamento da rede designada com o nome 'Sem Clave'.
- Rede de Detecção de Durações: 8 classes rotuladas representando valores rítmicos (semibreve a semifusa, com pontuação correspondente).

3.2. Arquiteturas utilizadas

Duas configurações distintas da família YOLO foram exploradas neste estudo, com o objetivo de avaliar seu desempenho na tarefa de Reconhecimento Óptico de Música.

Arquitetura YOLOv8 [8]:

- Backbone: Blocos C2f com módulo SPPF para extração multi-escala, reduzindo redundância computacional.
- Neck: C2f-PAN com conexões cruzadas bidirecionais.
- Head: Blocos CBS (Convolução-BatchNorm-SiLU) com ramos especializados.

Arquitetura YOLOv11 [9]:

- Backbone: Blocos C3k2 com atenção espacial C2PSA para discriminação de características locais.
- Neck: C3k2-PAN com seleção dinâmica de kernels.
- Head: CBS com adaptação dinâmica de receptores.
- Otimização: LAMB com taxa de aprendizado 0.001 e ajuste progressivo.

Tabela 1: Comparação entre as arquiteturas

Component	YOLOv8	YOLOv11
Backbone	C2f blocks	C3k2 with C2PSA
Neck	C2f-PAN	C3k2-PAN
Optimizer	AdamW	LAMB

3.3. Protocolo de Treinamento

Os modelos foram treinados em um subset de 137 partituras do dataset PRIUMS [10], que é composto de partituras monofônicas digitais (80% treino, 15% validação, 5% teste). Conforme mostrado na Figura 2, o banco de dados foi rotulado com respeito a cada uma das três redes. Uma com relação à claves e alturas, outro com relação às durações e outra com relação as linhas.



Figura 2: Exemplo da rotulagem realizada.

4. Resultados

A análise dos resultados obtidos com a utilização do YOLOv8 e YOLOv11 do pipeline indica ganhos relevantes quando o reconhecimento de claves é delegado ao LLM. A Tabela 2 apresenta as principais métricas comparando modelos treinados com e sem responsabilidade de detecção de claves, evidenciando o impacto positivo da integração com o Gemini 2.0 Flash nesse componente específico do OMR.

Tabela 2: Comparação de desempenho das configurações YOLOv8 e YOLOv11

Model	Precision	Recall
YOLOv8 (Com Clave)	0.653	0.587
YOLOv8 (Sem Clave)	0.7747	0.849
YOLOv11 (Com Clave)	0.712	0.623
YOLOv11 (Sem Clave)	0.8995	0.821

Os resultados revelam três descobertas principais.

- **Benefícios da Integração com LLM:** Observouse um ganho significativo de desempenho com a integração LLM no pipeline. Ambas as arquiteturas — YOLOv8 e YOLOv11 — apresentaram melhorias expressivas quando a tarefa de detecção de claves foi delegada ao LLM. O YOLOv8, por exemplo, registrou um aumento de 176,1
- **Desempenho superior do YOLOv11:** A arquitetura YOLOv11 superou consistentemente o YOLOv8 em todas as métricas e configurações avaliadas. Destaca-se, o valor obtido na mAP50 de

0,995 na configuração sem detecção de claves, sugerindo que os blocos C3k2 e os mecanismos de atenção espacial C2PSA são particularmente eficazes para tarefas de detecção de símbolos musicais. O valor de precisão de 0,8995 reforça ainda mais a acurácia do modelo na classificação dos elementos.

- **Compromisso precisão-recall:** A análise das métricas de precisão e recall indica um avanço relevante em precisão (acima de 95

Estes resultados têm importantes implicações para o campo. Os resultados obtidos corroboram a eficácia da abordagem híbrida LLM-YOLO para tarefas de OMR (Optical Music Recognition). O aumento de desempenho no mAP50-95 - uma métrica considerada robusta para avaliação de detecção - demonstra que a partição estratégica de tarefas entre modelos visão-linguagem e detectores especializados pode superar limitações tradicionais no reconhecimento de símbolos musicais. O mAP50 do YOLOv11 (0,995) destaca especialmente o potencial das arquiteturas modernas quando aplicadas em condições otimizadas, apontando para direções promissoras no desenvolvimento de sistemas de reconhecimento musical automatizado.

5. Conclusão

Este estudo propõe uma abordagem híbrida para Reconhecimento Óptico de Música (OMR), integrando modelos multimodais de linguagem (Gemini 2.0 Flash) com arquiteturas YOLO (YOLOv8 e YOLOv11) para transcrição de partituras manuscritas. Os resultados experimentais demonstram que a delegação da detecção de claves para um LLM melhora significativamente o desempenho da rede principal de detecção de objetos, permitindo que ela se concentre exclusivamente no reconhecimento de alturas e durações.

Implicações e Trabalhos Futuros: Os resultados obtidos indicam que LLMs multimodais podem complementar de forma eficaz sistemas OMR baseados em aprendizado profundo, ao assumirem o processamento do reconhecimento contextual de símbolos, enquanto detectores especializados focam na localização precisa e classificação dos demais elementos. Como trabalhos futuros, propõe-se uma extensão da abordagem para música polifônica [11] e também a inclusão de outros tipos de símbolos musicais a serem detectados por LLM's

A abordagem proposta apresenta potencial para contribuir com a preservação digital de acervos musicais históricos, combinando eficiência computacional a altos níveis precisão na transcrição automática.

Referências

- [1] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen. A survey on multimodal large language models. *arXiv pre-print arXiv:2306.13549*, 2023.
- [2] Y.-H. Cao, K. Ji, Z. Huang, C. Zheng, J. Liu, J. Wang, J. Chen, and M. Yang. Towards better vision-inspired

- vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13537–13547, 2024.
- [3] A. Pacha, J. Hajič Jr., and J. Calvo-Zaragoza. A baseline for general music object detection with deep learning. *Applied Sciences*, 8(9):1488, 2018.
 - [4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *arXiv preprint arXiv:1506.02640*, 2016.
 - [5] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2023.
 - [6] E. van der Wel and K. Ullrich. Optical music recognition with convolutional sequence-to-sequence models. *Zenodo*, 2017.
 - [7] L. Tuggener, I. Elezi, J. Schmidhuber, and T. Stadelmann. Deep watershed detector for music object recognition. *arXiv preprint arXiv:1805.10548*, 2018.
 - [8] D. Reis, J. Kupec, J. Hong, and A. Daoudi. Real-time flying object detection with yolov8. *arXiv preprint arXiv:2305.09972*, 2023.
 - [9] R.-Y. Ju and W. Cai. Fracture detection in pediatric wrist trauma x-ray images using yolov8 algorithm. *Scientific Reports*, 13:10375, 2023.
 - [10] J. Calvo-Zaragoza and D. Rizo. End-to-end neural optical music recognition of monophonic scores. *Applied Sciences*, 8(4):606, 2018.
 - [11] A. Ríos-Vila, J. Calvo-Zaragoza, and T. Paquet. Sheet music transformer: End-to-end optical music recognition beyond monophonic transcription. *Journal of New Music Research*, 2024.