EFFICIENT NOTATION ASSEMBLY IN OPTICAL MUSIC RECOGNITION

Carlos Penarrubia[1], Carlos Garrido-Munoz[1], Jose J. Valero-Mas[2], Jorge Calvo-Zaragoza[1]
[1]U. I. for Computing Research, University of Alicante, Spain
[1]{carlos.penarrubia, carlos.garrido, jorge.calvo}@ua.es
[2]Music Technology Group, Universitat Pompeu Fabra, Spain
[2]josejavier.valero@upf.edu

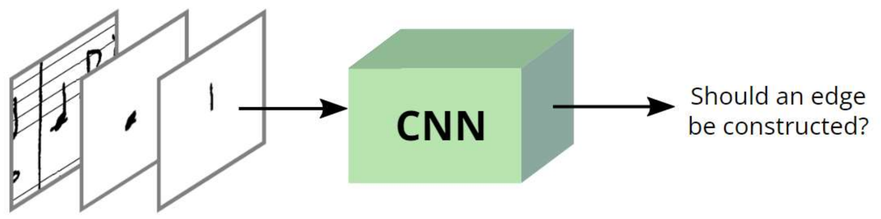
INTRODUCTION AND RELATED WORKS

TRADITIONAL STEPS IN OMR

1. **Image preprocessing**: tasks such as binarization, distortion correction, or stave separation.
2. **Music symbol detection**: bounding box detection and classification
3. **Notation assembly**: where the independent components are related to each other to reconstruct the musical notation.
4. **Encoding**: in which the recognized notation is exported to a specific language that can be stored and further processed by computational means

Related works

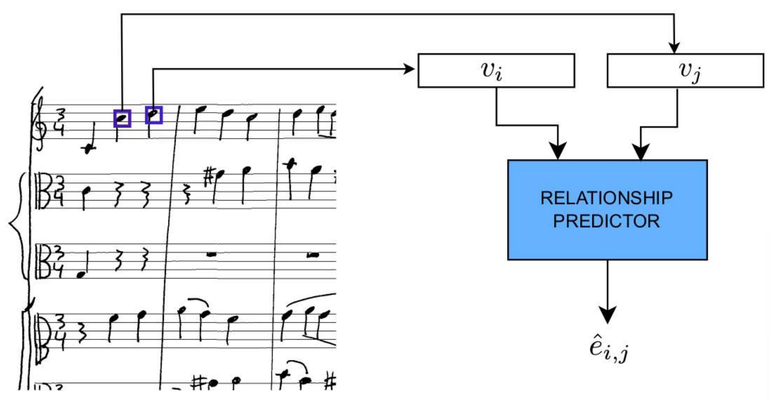
- Only **one existing work** focused on the relationship retrieval: Pacha et al [1].
- Based on a Convolutional Neural Network (CNN)
- **Tremendously inefficient**: requires the independent construction and classification of an image for each pair of nodes



EXPERIMENTAL METHODOLOGY

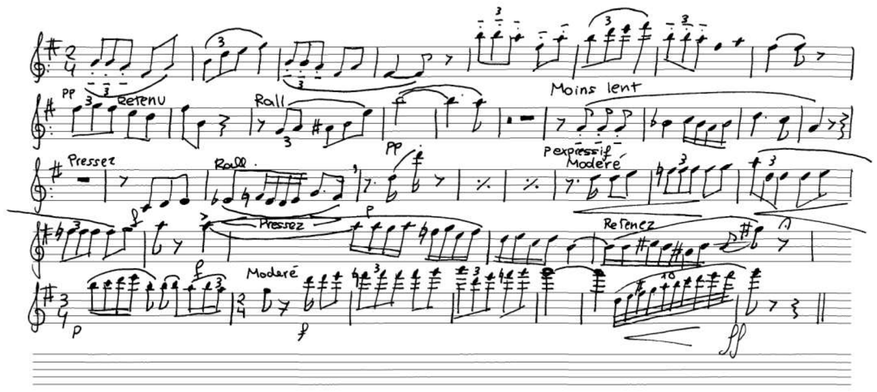
PROBLEM FORMULATION

The notation assembly stage can be seen as a **relationship predictor** of a graph where the nodes are the music symbols and the edges are the relations between the nodes.



DATASET: MUSCIMA++

5 fold cross validation



RELATIONSHIP PREDICTION PROPOSALS

Node: 20-dimensional feature vector:
- Bounding box: top-left and bottom-right normalized values
- Class information: 16-dimension learnable embedding layer

Node feature vector

Top	Left	Bottom	Right	Class information (16-dimension)

- **$MLP_{64,512}$**: A three-layered fully-connected network comprising two hidden layers with 64 and 512, respectively, with Rectifier Linear Unit (ReLU) activations and a single output unit to compute the score of the binary classification.
- **MLP_{32}**: A two-layered fully-connected network comprising a 32-unit hidden layer and ReLu activation and a single unit as output
- **Asymmetric kernels**: are implemented as two different 4-layered MLP comprising 512, 1024, 512, and 256 units, respectively, with ReLU activation. The idea is to generate two 256-dimensional embeddings—two points in different Hilbert spaces—to then compute the similarity through the dot product.

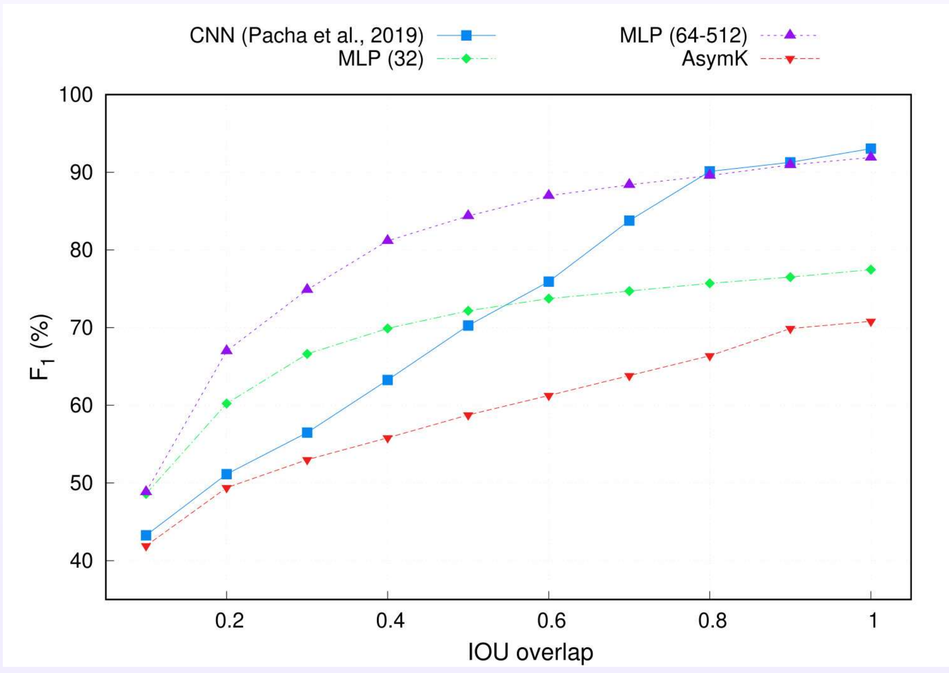
RESULTS AND DISCUSSION

EFFICIENCY

Execution time (ms)	AsymK	MLP_{32}	$MLP_{64,512}$	CNN
	<0.5	55	176	$>1.5 \cdot 10^6$

Efficiency results in terms of the per-page absolute execution time (in milliseconds) on the MUSCIMA++ corpus for the different notation assembly methods assessed. Each value corresponds to the average execution time obtained with **10 different iterations** over all test samples.

EFFICACY



CONCLUSIONS

- As efficiency and efficacy are opposite criteria to be optimized, there is not a predictor that performs the best result in both aspects.
- We can justify that the complex MLP is the model that gets an efficacy very close to the best model (the CNN) while being several orders of magnitude faster and even outperforming the CNN when the IoU is degraded
- AsymK is the most efficient solution.

BIBLIOGRAPHY

[1] A. Pacha, J. Calvo-Zaragoza, and J. Hajic Jr., "Learning notation graph construction for full-pipeline optical music recognition," in Proceedings of the 20th International Society for Music Information Retrieval Conference, 2019, pp. 75–82.

Acknowledgements: Work produced with the support of a 2021 Leonardo Grant for Researchers and Cultural Creators, BBVA Foundation. The Foundation takes no responsibility for the opinions, statements and contents of this project, which are entirely the responsibility of its authors.