

Incremental Supervised Staff Detection

Alexander Pacha

Institute of Information Systems Engineering

TU Wien, Austria

alexander.pacha@tuwien.ac.at

Abstract—Music scores written in modern notation use staves as a reference system for assigning semantics to the individual symbols that appear in the score. Detecting this structural element is, therefore, a natural step in most Optical Music Recognition systems. However, many systems struggle to reliably detect staves. This paper investigates whether computers can learn to detect staves with a convolutional neural network given only a small set of images for which annotation are available. After an initial training phase, the network is asked to make prediction on a larger test dataset. A human annotator reviews the predictions and approves or rejects samples. Approved samples will be added to the training set for the next iteration to incrementally expand the training set and allow the network to operate well on a variety of music scores.

After four iterations, we were able to obtain staff bounding box annotations for 14,000 out of 20,000 scores in our dataset. Although the evaluated approach has structural flaws that lead to imprecise results and deficits when detecting non-straight staves, it can serve as a viable starting point for future staff detection systems.

Index Terms—Optical Music Recognition, Music Staff, Object Detection

I. INTRODUCTION

Music scores written in modern notation use staves - typically five parallel lines - as a reference system for notes. Staves are further divided into individual measures by bar lines to provide visual guidance for the reader. Detecting these structural elements is of fundamental importance to virtually every Optical Music Recognition (OMR) system. Most systems use staff detection as a preprocessing step to break the image down into meaningful sub-regions that can be processed individually [1]. A variety of methods have been proposed to robustly detect staves, or more precisely, individual staff lines before removing them [2]. This was a prerequisite for techniques such as connected-component analysis in the symbol detection stage. However, in the last few years, new approaches were proposed that are capable of detecting music objects without the need for staff removal [3], [4], but still require segmentation of the image into individual staves to overcome computational restrictions.

If measures can robustly be detected in music scores it is only a small step towards doing the same with staves. A universal staff detector in combination with a robust measure detector provides a vital framework for subsequent steps and is by itself already a useful tool, allowing users to quickly navigate through scores. One could also envision a collaborative tool where humans manually transcribe a music piece measure by measure, filling in an empty scaffolding of the score

that was automatically generated. Such a tool could facilitate the crowd-sourced transcription of a large body of music, like the IMSLP [5]. Modern machine learning approaches with convolutional neural networks can reliably solve object detection tasks given a sufficient amount of annotated data. At the same time, they promise that the trained model will generalize to new data if the training data contains a sufficient amount of variation. To put this claim to the test and to avoid manually labeling thousands of images, this paper reports on experiments for training a staff detector with limited ground truth data.

II. RELATED WORK

Under ideal conditions, staff lines in music scores are straight, parallel lines than span a large portion of the image. A simple, yet effective way to detect those lines is to perform a projection along the x-axis and look for very high peaks [6]. But this method deteriorates quickly if the image is slightly rotated or exhibits other distortions (see Fig. 1).

More robust methods were developed in the last 30 years including scan lines, Hough transformations and stable paths [7], [8]. The most notable effort on comparing these methods was the ICDAR Music Scores Competition [9]–[11], which produced the CVC-MUSCIMA dataset [12] and showed that several algorithms operate well even at high levels of synthetic degradation. So, is staff detection a solved problem? It is genuinely hard to answer because music scores can be extremely diverse. Even if an algorithm works well on a certain dataset, claiming it solves staff detection would require a thorough evaluation on a very large dataset such as the IMSLP [5] which contains nearly 500,000 scores and over 10 million pages. To the best of our knowledge, no one ever attempted such an evaluation. A recent test of commercial OMR applications indicates that many systems already struggle at this early stage if the input score is just 2° rotated [13], leaving plenty of room for improvement.

In recent years, the research interest shifted to machine learning methods which use convolutional neural networks to detect and remove staff lines [14]–[16]. Similarly, the detection of measures was also recently attempted with machine learning methods [17] and shows promising results. The main drawbacks of these machine learning methods are that they require a large amount of annotated data and while they do operate well within the boundaries of what data they saw during the training, it is not guaranteed that they work satisfactorily on new data.

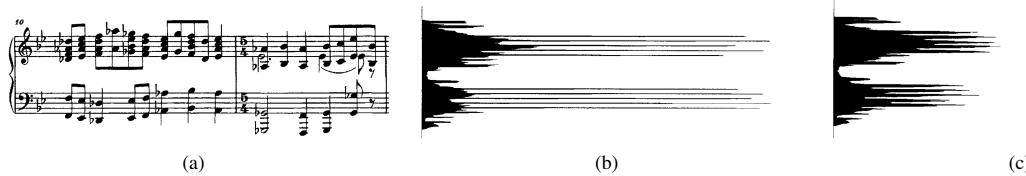


Fig. 1. Primitive staff detection by projection from Bainbridge et al. The input score (a) with its projection profile under ideal conditions (b) and its projection profile after slightly rotating the image which reflects more realistic conditions (c).

III. SELF-LEARNING STAFF DETECTION

Ideally, we would like to have a versatile staff detector that can robustly detect staff lines under various conditions even at the presence of distortions. A state-of-the-art object detector [18] can robustly detect objects in images. Simple objects like staves and measures are no exception. It feels natural to further examine this avenue. However, due to the lack of a large annotated dataset (and out of sheer curiosity), an interactive approach is devised to build that staff detector. The idea is that you start with a small amount of annotated data, then you train on that data and perform inference on a much larger dataset. The detector will hopefully produce correct results for at least a few previously unseen pages. A human reviewer then examines the results. Those images with correctly detected staves are added to the set of scores for which we have correct annotations and used as training data for the next iteration. In theory, this process can be repeated until we know the staff positions of every music score in our database. Figure 2 illustrates this procedure.

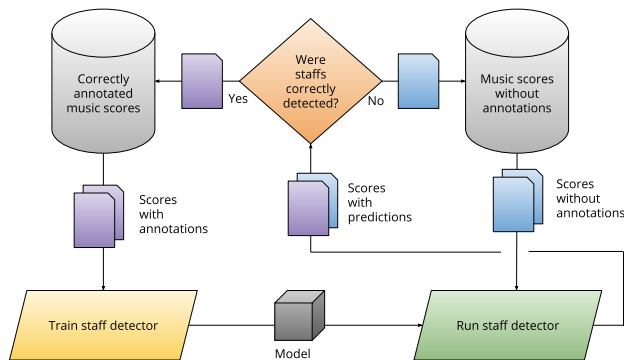


Fig. 2. The iterative workflow to train the self-learning staff detector.

For detecting bounding boxes, a Faster R-CNN [18] detector was used. The training regime is equivalent to the measure detector as described by Waloschek et al. [17] and the source code is publicly available¹.

The initial dataset for which ground truth annotations exist is the MUSCIMA++ dataset [19]. The target dataset for which no annotations were available is a large collection of multiple datasets containing approximately 20,000 images, ranging

from typeset, born-digital images taken from the DeepScores dataset [20] to manuscripts with substantial degradation from the IMSLP. The curriculum for the first iteration was to generalize from the initial 100 pages to the entire CVC-MUSCIMA dataset of 10,000 images. After that first warm-up round, consecutive iterations evaluated the entire dataset for which the correct bounding boxes are not yet known.

IV. RESULTS

After four iterations, we obtained reasonable bounding box annotations for 14,000 out of 20,000 scores. For the remaining 6,000 images, the detector was not able to produce acceptable predictions. While 14,000 seems high, it should be noted that the first 10,000 were just the CVC-MUSCIMA dataset for which correct predictions were easily obtained during the first iteration. Each consecutive iteration yielded a higher number of correct results. However, the idea that simply continuing for another few iterations will produce accurate results for the entire dataset, unfortunately, turned out to be unfeasible because of two reasons:

- The used detector operates on rectangular regions only. Real scores, however, are often bent or rotated, therefore a larger target bounding box is needed to cover the rotated staff. The model then learns to also use this larger bounding box for staves that are not rotated, causing increasingly inaccurate predictions.
- Prediction are rarely as precise as if the bounding boxes were annotated manually. If the annotator accepts a sample with such slight imperfections, he introduces a small error into the training set. These errors accumulate from iteration to iteration, leading to increasingly inaccurate results. This problem can best be observed with staves that should be smaller, because they are preceded by instrument names that we do not consider to belong to the staff. Therefore, these staves should be smaller. Given the lack of specific training data that exhibit this property, the detected regions frequently exceed the actual staff and do include the instrument name (see Fig. 3).

Reviewing the 14,000 images for which the neural network produced acceptable bounding boxes revealed that most of these images contain straight staves. Unsurprisingly, many of the bounding boxes exceed the boundaries of the contained staff quite significantly. Manuscripts with skewed staves were the hardest to predict correctly. A few examples of the results are given in the Appendix.

¹<https://github.com/OMR-Research/MeasureDetector/tree/master/StaffDetector>

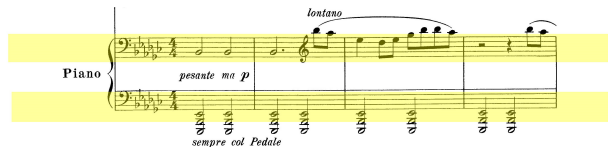


Fig. 3. A staff that is preceded by an instrument name with its predicted bounding boxes shown as transparent yellow overlay.

V. DISCUSSION AND CONCLUSION

Reviewing images whether they contain appropriate bounding box predictions can be done much faster than actually annotating each image individually. Checking an individual image took less than a second, making this process very efficient. While the process is laborious at the beginning where less than 100 out of 1000 images can be moved from the pool of images without annotations to the pool with annotations, this procedure eventually gets even more efficient, as fewer images have to be reviewed each iteration. However, very dense scores, bent scores, or rotated scores still pose a major challenge to the detection model. The generalizability of the described approach to these scores remains limited.

The downside of this workflow is the decreasing quality of the bounding box predictions. In some scenarios, it can be acceptable to only have a coarse approximation, e.g., for cropping the image into sub-regions that are further processed, but it might not be suitable in other cases, e.g., if the staff line distance is estimated from the bounding box. Finally, in future work it would be interesting to compare this method to similar methods like Mask R-CNN [21] that perform instance segmentation and promise to also work on more irregular shapes. To allow for a better comparison, the best model will be made publicly available on Github. Large parts of the dataset can also be shared upon request.

REFERENCES

- [1] J. Calvo-Zaragoza, J. Hajić jr., and A. Pacha, "Understanding optical music recognition," *Computing Research Repository*, 2019. [Online]. Available: <https://arxiv.org/abs/1908.03608>
- [2] A. Rebelo, I. Fujinaga, F. Paszkiewicz, A. R. Marcal, C. Guedes, and J. d. S. Cardoso, "Optical music recognition: state-of-the-art and open issues," *International Journal of Multimedia Information Retrieval*, vol. 1, no. 3, pp. 173–190, 2012.
- [3] A. Pacha and H. Eidenberger, "Towards a universal music symbol classifier," in *14th International Conference on Document Analysis and Recognition*, IAPR TC10 (Technical Committee on Graphics Recognition). Kyoto, Japan: IEEE Computer Society, 2017, pp. 35–36.
- [4] A. Pacha, J. Hajić jr., and J. Calvo-Zaragoza, "A baseline for general music object detection with deep learning," *Applied Sciences*, vol. 8, no. 9, pp. 1488–1508, 2018. [Online]. Available: <http://www.mdpi.com/2076-3417/8/9/1488>
- [5] Project Petrucci LLC, "International music score library project," <http://imslp.org>, 2006.
- [6] D. Bainbridge and T. Bell, "The challenge of optical music recognition," *Computers and the Humanities*, vol. 35, no. 2, pp. 95–121, 2001.
- [7] C. Dalitz, M. Droettboom, B. Pranzas, and I. Fujinaga, "A comparative study of staff removal algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, pp. 753–766, 2008.
- [8] J. d. S. Cardoso, A. Capela, A. Rebelo, C. Guedes, and J. Pinto da Costa, "Staff detection with stable paths," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 6, pp. 1134–1139, 2009.

- [9] A. Fornés, A. Dutta, A. Gordo, and J. Lladós, "The ICDAR 2011 music scores competition: Staff removal and writer identification," in *International Conference on Document Analysis and Recognition*, 2011, pp. 1511–1515.
- [10] A. Fornés, A. Dutta, A. Gordo, and J. Lladós, "The 2012 music scores competitions: Staff removal and writer identification," in *Graphics Recognition. New Trends and Challenges*, Y.-B. Kwon and J.-M. Ogier, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 173–186.
- [11] A. Fornés, V. C. Kieu, M. Visani, N. Journet, and A. Dutta, "The IC-DAR/GREC 2013 music scores competition: Staff removal," in *Graphics Recognition. Current Trends and Challenges*, B. Lamiroy and J.-M. Ogier, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 207–220.
- [12] A. Fornés, A. Dutta, A. Gordo, and J. Lladós, "CVC-MUSCIMA: A ground-truth of handwritten music score images for writer identification and staff removal," *International Journal on Document Analysis and Recognition*, vol. 15, no. 3, pp. 243–251, 2012.
- [13] J. Noll, "Intelligentes notenlesen," *c't*, vol. 18, pp. 122–126, 2019.
- [14] J. Calvo-Zaragoza, L. Micó, and J. Oncina, "Music staff removal with supervised pixel classification," *International Journal on Document Analysis and Recognition*, vol. 19, no. 3, pp. 211–219, 2016.
- [15] J. Calvo-Zaragoza, A. Pertusa, and J. Oncina, "Staff-line detection and removal using a convolutional neural network," *Machine Vision and Applications*, pp. 1–10, 2017.
- [16] A.-J. Gallego and J. Calvo-Zaragoza, "Staff-line removal with selectional auto-encoders," *Expert Systems with Applications*, vol. 89, pp. 138–148, 2017.
- [17] S. Waloschek, A. Hadjakos, and A. Pacha, "Identification and cross-document alignment of measures in music score images," in *20th International Society for Music Information Retrieval Conference*, 2019.
- [18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems* 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 91–99.
- [19] J. Hajić jr. and P. Pecina, "The MUSCIMA++ dataset for handwritten optical music recognition," in *14th International Conference on Document Analysis and Recognition*, Kyoto, Japan, 2017, pp. 39–46.
- [20] L. Tuggener, I. Elezi, J. Schmidhuber, M. Pelillo, and T. Stadelmann, "Deepscores - a dataset for segmentation, detection and classification of tiny objects," in *24th International Conference on Pattern Recognition*, Beijing, China, 2018.
- [21] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

APPENDIX

The following appendix contains a small selection of scores to demonstrate the variety of the data and the performance of the detector. The predicted bounding boxes are superimposed as transparent yellow overlay.

DER ENGEL WANDERUNG.

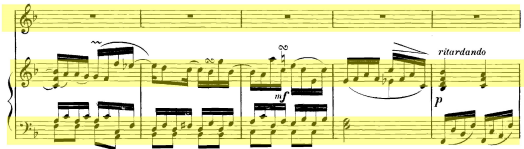
Dichtung von Emil Tscsano.

Musik von Anthony Philip Heinrich of New York

Piacevole.

Gesang. 

Piano. 



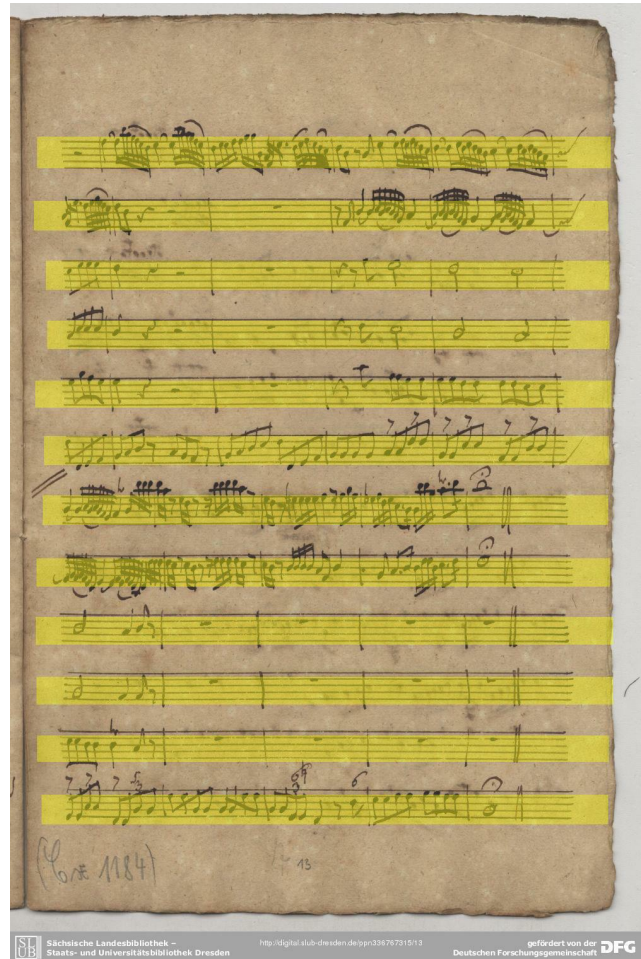
Quasi Allegretto, con Grazia.

Es glebt viel Engels Kin - der, Im wei - ten Him - mels - zelt; Die



36 

40 



Sächsische Landesbibliothek –
Staats- und Universitätsbibliothek Dresden

<http://digital.stuk-dresden.de/pr/33678731513>

gefordert von der
Deutschen Forschungsgemeinschaft DFG

