The DeepScoresV2 Dataset and Benchmark for Music Object Detection

Lukas Tuggener
ZHAW Datalab & USI
tugg@zhaw.ch

Yvan Putra Satyawan
ZHAW Datalab
orcid.org/0000-0002-6375-8308

Alexander Pacha
Vienna, Austria
alexander.pacha@tuwien.ac.at

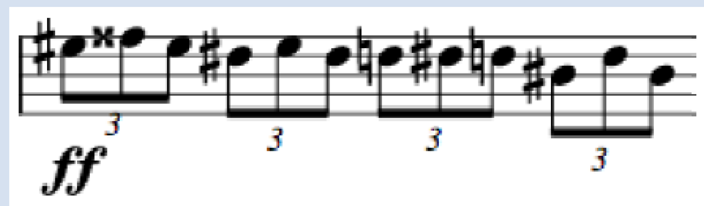
Jürgen Schmidhuber
IDSIA & USI
juergen@idsia.ch

Thilo Stadelmann
ZHAW Datalab
stdm@zhaw.ch

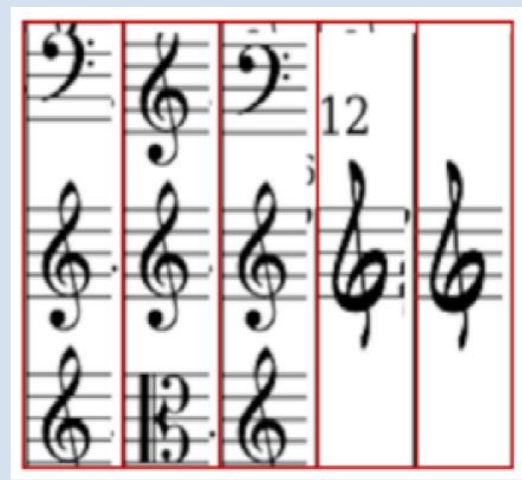
What is *DeepScores* ?

DeepScores is a vast collection containing 300‘000 sheets of written music. The music has is rendered in high quality and annotated with ground truth for object classification, object detection and sematic segmentation. While *DeepScores* has been designed with Optical Music Recognition in mind does it pose a number of very interesting challenges for general computer vision. Among them are:

- Extremely imbalanced classes
- High size disparity between classes
- Large number of symbols per page
- Class label can depend on context

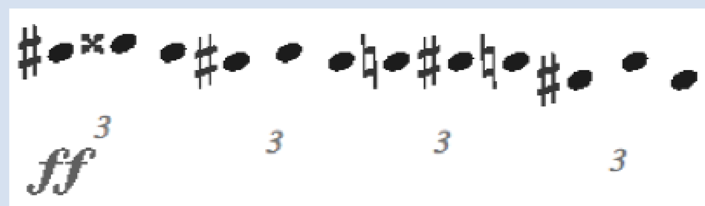


Example data



Classification examples

Bounding box annotation

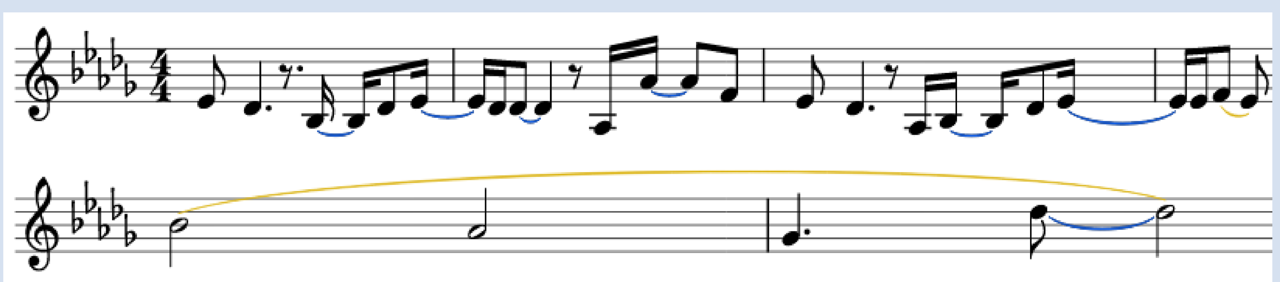


Pixel level annotation

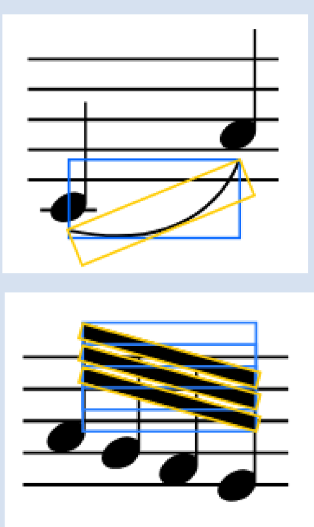
Improvements in DeepScoresV2

Additional Annotated Classes

DeepScoresV2 features a wider variety of annotated symbol classes, increasing the total number of annotated classes to 135. Notably does it also include non-fixed shape symbols such as ties (blue) and slurs (yellow) as seen below.



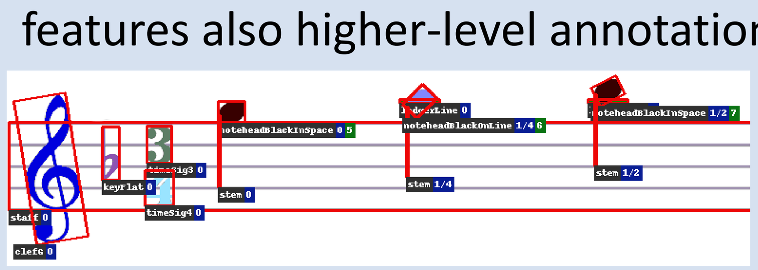
Oriented Bounding Boxes



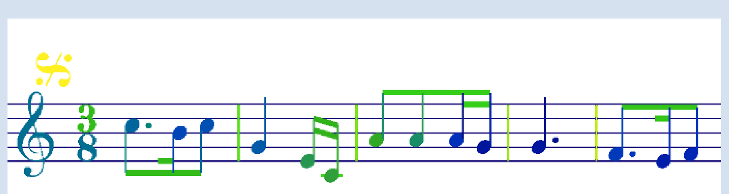
Annotations in object detection datasets are usually stored as rectangles which are aligned with the base image (blue). This can lead to very inaccurate annotations and overlaps, especially for long and thin objects that stand at an angle.
We incorporated bounding boxes at an angle (yellow) into DeepScoresV2 to address this issue.

Rhythm and Pitch Information

Optical music recognition is a challenging task beyond the mere localization and identification of notation elements. DeepscoresV2 features also higher-level annotations, namely the onset (starting beat) of every object and the pitch of the noteheads.



Instance Segmentation Annotations

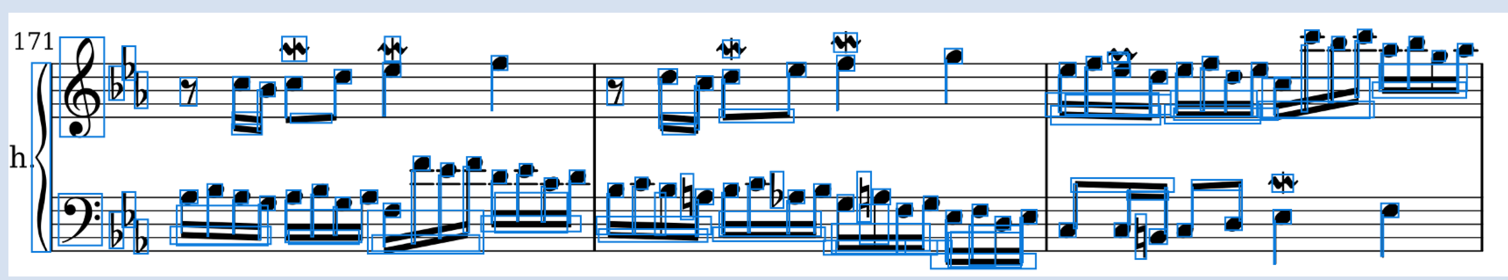


To allow researchers the maximum freedom in their model design have we also added instance segmentation ground truth.

Baseline Results

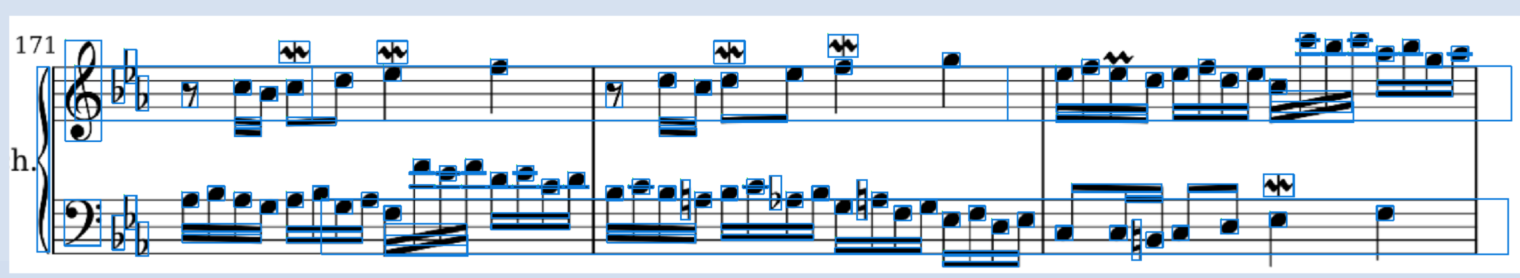
Deep Watershed Detector

The Deep Watershed Detector[1] was trained on full resolution images, with ledgers and staffs disabled due to overlap with other symbols. The biggest issues of the DWD in this application is the low accuracy of the bonding boxes (see predictions below) which leads to an AP@0.5 of *0.503* but a mAP of only *0.203*



Faster R-CNN

We used Faster R-CNN with a HRNet backbone[2] as our second baseline. It achieves a very high scores with an AP@0.5 of *0.799* and a mAP of *0.700*. It produces a very high bounding box accuracy (as seen below); however, the stems are missed completely, and its performance degrades significantly for rare symbols.



Ressources

The dataset, instructions as well as all the code needed to test the dataset and recreate the presented baselines is available at:
https://zenodo.org/record/4012193
https://github.com/tuggeluk/DeepWatershedDetection/tree/dwd_old
https://github.com/tuggeluk/mmdetection/tree/DSV2_Baseline_FasterRCNN