TRANSCRIBING MANUSCRIPTS IN MENSURAL NOTATION : A COMPREHENSIVE PIPELINE

Yoeri Uyttendaele
ESAT-STADIUS KU Leuven
yoeri.uyttendaele@kuleuven.be

Bart De Moor
ESAT-STADIUS KU Leuven
bart.demoor@kuleuven.be

ABSTRACT

We present a complete OMR pipeline for transcribing music in mensural notation. Given a scanned handwritten or printed manuscript, our model produces a MusicXML or MEI file, from which the user can preserve the mensural notation or transcribe to Common Music Notation (CMN). In comparison with other works, our pipeline achieves state-of-the art results. As our second contribution, we introduce a dataset written in mensural notation comprising annotated folia from illuminated choir books, printed works and chansonniers. This dataset aims to be representative for polyphonic renaissance music written in monophonic notation: folia span the whole of the renaissance period, contain several hand writing styles and show various degrees of image degradation. On this dataset and the SEILS dataset, our OMR model achieves state-of-the art results. Our pipeline builds on an ensemble of YOLOR object detection models which aim to mitigate class imbalance and the problem of small symbol detection. On the other hand, we construct a novel staff line parsing algorithm which detects all lines on the constituting staff. Our algorithm is capable of consistently detecting even severely degraded staff lines, and is equally applicable to the domain of music written in polyphonic notation. With our OMR pipeline, we facilitate access to music written in mensural notation for early music researchers, musicians and their audiences. With our dataset we aim to fill the gap of a versatile benchmark dataset for mensural notation. In doing so, we provide a means to make validation of OMR research on mensural notation even more robust.

1. INTRODUCTION

To aid musicological research, libraries around the world have been digitizing their collections of music manuscripts extensively [1, 2]. However, creating a digital scan of a music manuscript is but a first step in retrieving its musical content. The field of Optical Music Recognition (OMR) aims to extract the musical information from a scan of a musical score and write this information to a machine-readable format.

In this paper we present a full pipeline OMR for music written in mensural notation: provided a scan of a printed or hand-written score, we detect the music symbols and write the music information to an MEI or **mens file in mensural notation. The user can opt to transcribe the music to CMN which can then be written to MEI or MusicXML from which one can produce the associated audio waveform. After the object detection step and pitch classification step, the user has the option to make corrections to the detected objects and pitches. With the aid of *Mensural-score-to-CMN-score* – i.e., a toolkit to translate mensural MEI files into common music notation – a transcription to Common Music Notation is possible [3]. To the best of our knowledge, this is the first full pipeline applicable to the whole of western Renaissance polyphony, a wide variety of handwriting styles and degraded documents, including books such as the recently discovered Leuven Chansonnier [1]. Moreover, we demonstrate techniques to overcome previously existent and general OMR challenges such as class imbalance and small symbol detection. Additionally, we introduce a new staff detection paradigm surmounting previous staff line retrieval issues on degraded staffs [4].

The remainder of this paper is organized as follows: Section 2 describes the state of optical music recognition as applied to early music while Section 3 outlines our dataset. In Section 4 we propose and detail our OMR pipeline. Section 5 presents the results of our proposed method on the test data. Lastly, Section 6 contains a brief conclusion of our work.

2. RELATED WORK

The ultimate goal in the field of optical music recognition is to create a fully automated system that can interpret written or printed music in a digital image. Unfortunately, only a handful of OMR systems exist for early music manuscripts. Perhaps the most well-known is *Aruspix* [5], Aruspix produces a musical score based solely on a manuscript scan, and comes with a graphical user interface in which the user can edit the result alongside the original manuscript. Torras et al. aim to alleviate the amount of editing by using a language model [6]. Pugin et al. have compared Aruspix to another system named *Gamut*, but concluded that Gamut is inferior to Aruspix, as it nei-



© F. Author, S. Author, and T. Author. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** F. Author, S. Author, and T. Author, "Transcribing manuscripts in mensural notation : A comprehensive pipeline", in *Proc. of the 23rd Int. Society for Music Information Retrieval Conf.,* Bengaluru, India, 2022.

ther produces a musical score, nor does it come pre-trained for early music [7]. Huang et al. produce a pipeline for early music notation, however they overfit their model on a single handwriting style and deploy a staff parsing algorithm which displays difficulty on degraded staffs [8]. The work of Huang et al. builds upon the *classical* OMR system architecture, as specified by Rebelo et al. [9]. This architecture consists of multiple subsequent steps, which in different works of research are modified, or merged. For example, in 2018, Hajič and Pecina have shown how a convolutional neural network can perform the *isolating primitive elements* and the *symbol classification* steps [9] of the OMR pipeline, but their method still requires a separate *binarization* step [10]. Although binarization is indeed an important step in the OMR pipeline by Rebelo et al., it is far more common for CNN's to work with (the original) full-color images: using full color images as input instead of grayscale images provides the network with more information. Castellanos et al. have shown that fully convolutional neural networks can separate foreground from background in full-color manuscript scans [11].

For early music manuscripts, separating foreground from background information is a particularly pertinent and cumbersome task: music symbols can be faded, or bleed-through from the other side of the page. Several authors have proposed methods on how to handle these issues, including Burgoyne et al. [12], who compare and combine models which take into account the image on the backside of the folium, and models that don't. Furthermore, the output of the binarization step can be used to locate the region of interest on musical imagery, even when the folium's 3D orientation in a photograph needs to be taken into account [13]. In the same work, it is made clear how the location of staves can be inferred from the binarized image by using histogram analysis. In this case, the staves are located based on the location of *any* musical content, rather than looking for the staves themselves.

Pacha and Calvo-Zaragoza attempt the entire music symbol recognition step using a single model for early music [14]. The authors use a second convolutional neural network to classify the position of a detected note on the staff, in order to determine its pitch. Although Pacha et al. focus on a specific mensural writing style in their work, the underlying characteristics of the dataset are similar to those of this work: the authors attempt to detect a large number of classes (32) using the same model, while the prevalance of these classes is highly imbalanced. This work too is trained on but a single corpus. Pacha and Calvo-Zaragoza mention that for the least prevalent classes – "dots, barlines, or all types of rests" – the recall drops heavily. Huang et al. also mention issues with dot symbols and attribute this to class imbalance in the training dataset [8].

3. DATASET

We introduce and publish a dataset designed to fill the gap of representative datasets for renaissance music written in mensural notation. Our dataset contains annotations to scans of illuminated folia from choirbooks, madrigals from printed works and folia from chansonniers. In total we present 250 annotated folia spread across fourteen books. The images show a variety of handwriting styles at different levels of degradation and ink bleed-through.

The fourteen manuscripts encompass folia from six illuminated choirbooks *Vienna, Österreichische Nationalbibliothek, Musiksammlung Mus. Hs. (A-Wn) 15495* , *Brussels, Koninklijke Bibliotheek van België / Bibliothèque royale de Belgique Ms. 228 and Ms. IV.922*, *Munich, Bayerische Staatsbibliothek D-Mbs Mus. Ms. F*, *'s-Hertogenbosch, Illustre Lieve Vrouwe Broederschap Ms. 72A*, and *Città del Vaticano, Biblioteca Apostolica Vaticana Ms. Chigi C VIII 234* [1],
five chansonniers: (*Dijon, Bibliothèque Municipale Ms. 517 Wolfenbüttel, Herzog August Bibliothek, Codex Guelf. 287 Extravag. Copenhagen, The Royal Library, MS Thott 291* 8° , *Nivelle, Paris, Bibliothèque nationale, Rés. Vmc. ms. 57* [2], *Leuven - Alamire Foundation B-AF Ms. 1* [1]), and three printed works: *Madregali ariosi a quatro voci* by Antonio Gardano, the *Compieta, falsi bordoni, mottetti, et litanie della Madonna* by Giulio Belli and *Ghirlanda Sacra* [15].

We let an expert in the field of early music notation manually label 60 pages from these manuscripts. Next, we train a preliminary convolutional neural network on these labeled pages. This trained model labels another 190 pages, subsequently the expert corrects the errors to obtain a dataset of 250 pages and 80165 symbols. This data is published along with this work, with each staff annotated separately in the CSV, **mens and MEI formats.

The folia contain 37 classes which generally fall into 8 categories, namely: *notes* (maxima,longa, breve, semibreve, minim, semiminim, fusa, semifusa), *colored notes* (maxima,longa, breve, semibreve), *rests* (maxima,longa, breve, semibreve, minim, fusa), *clefs* (c clef, f clef, g clef), *accidentals* (flat, sharp), *mensurations* (Modus (im)perfectum tempus, prolatio (im)perfectum major/minor, tactus brevis/semibrevis), *digits* (1, 2, 3), and *others* (repeat, barline, fermata, dot, custos).

In the ground truth annotations we do not explicitly annotate obliqa's or any other type of ligature note, instead, we parse the ligature following Apels heuristic [16]. We do not interpret the written notes, e.g. we do not change duration for coloration or distinguish dots of division or augmentation. Other than ligature notes, music symbols are annotated solely based on their geometry.

4. METHODOLOGY

In this section we broadly sketch the OMR pipeline. This pipeline differs from the classical one as proposed by Rebelo et al. [9]: the border removal step is now largely superfluous given the staff detection by neural network, while the binarization step is absolved by the symbol detection neural network. Indeed, removing the staff may lead to the

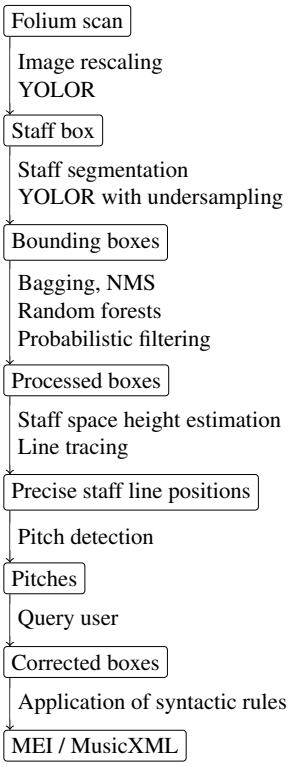

Figure 1: Starting from a folium scan, we obtain a transcribed music document. Some traditional preprocessing tasks are absorbed by the first-pass neural network. Staffs margins are inferred from the object detection output. Specialist models handle the detection of rare symbols.

unwanted removal of music symbols, therefore we choose to keep the staff intact. The staff is detected by a state-of-the-art neural network, YOLOR [17]. Subsequently, we segment the found staff and apply specialized models to determine the symbol class. The segmentation allows us to overcome the small size of symbols such as *dot* or *fusa rest*. Moreover these neural nets are trained by undersampling the most frequent classes and using super classes for the rarest of symbols. Probabilistic filtering does away with low confident false positive staffs and false positive symbols. A Dynamic programming approach infers the staff line position after which a CovNet assigns a pitch to the music symbols when applicable. Next, we apply the ligature parsing algorithm by Apel [16]. Finally, we write away the found symbols to several formats such as XML and MEI upon which corrective steps can be made by uploading the XML file to an annotation program [18]. The resulting pipeline is visualized in Figure 1.

4.1 Preprocessing

For the purpose of detecting the staffs, we rescale the input image to 1024×1024 pixels. For symbol detection on the staff we preprocess by segmenting the staff in 4 overlapping parts each of which is resized to 640×640 pixels. This rescaling allows us to detect both small and large symbols during inference. We choose not to implement automated border cropping techniques or staff removal techniques, as these – much like staff removal techniques – may lead to the removal of true positive symbols [19]. For this same reason, we abstain from binarization prior to object detection; to distinguish foreground from background, we let the CNN learn from degraded images, thereby alleviating us from the need of a binarization step [20].

4.2 Staff detection and music symbol detection

Scans or pictures may contain various forms of noise surrounding the folium of interest such as rulers, color checks or parts of adjacent pages. When present, the size of the musical objects in the frame is drastically reduced. Moreover, the symbol size within a class may also vary – in addition to the intrinsic size differences between classes. A model which detects the outline of the staff after which symbols can be detected, alleviates aforementioned problems. We choose the YOLO architecture as it is designed to facilitate the recognition of objects at different scales, an advantage over models with fixed anchor box sizes. Specifically, we use transfer learning on a YOLOR-D6 model that is pre-trained on the MS COCO dataset on which YOLOR-D6 achieves state-of-the-art results [17, 21]. We deploy this model for staff detection as well as symbol detection.

The training dataset for these models is created by randomly sampling 197 images from ten out of fourteen books. Furthermore, we add 3 pages from the *Missa Prolationum* by Ockeghem in order to have samples for perfect major prolation, thereby ensuring representation for every class in the training data. The training dataset consists of 64591 symbols in total. The other 50 pages in the ground truth data are spread over the entire collection of 14 books, two books of which are exclusively purposed for testing, the *Dijon Chansonnier* and the *Wolfenbüttel Chansonnier*, while two books are used solely for validation, the *A-Wn 15495* and the *Ghirlanda sacra*. The test set contains 35 pages, the validation set 15 pages.

To enhance robustness of the staff detection model at various scales, we augment the image by randomly rescaling and randomly positioning the folium scan on a 1024×1024 pixels background. We augment our data set of 200 pages to obtain 1000 images for staff detection. The scaling ranges from 600×600 pixels to the full background size. Alternatively, the YOLOR model allows for the use of mosaic image augmentation to improve detection at different scales, however, we find this leads to the false positive detection of parts of staffs on the adjacent page. The model is trained using the Adam optimizer and weight decay. We use early stopping regularization and find that the model achieves optimal performance after training for 41 epochs. We grid search the training set to find that a confidence level of 25 $\%$ results in the least amount of false negatives. We discuss the removal of false positives later on.

As discussed previously, music notation has inherent

class imbalance. Moreover, symbols such as dots or some rests are small. We mitigate this problem by performing object detection on a segmentation of each staff with specialized models. This way symbols take up a larger proportion of the CNN's input image and become more easily detectable. To attenuate class imbalance, we segment each staff in the training dataset into 4 overlapping parts. The overlap ensures every symbol is seen by the neural net. The beginning of a staff generally contains rare symbols such as clefs, mensuration signs and accidentals. Therefore, we build a separate model for the leftmost part of the staff, and train it only for windows of leftmost parts of staves. This helps combat the class imbalance as it increases the relative occurrence of these rare symbols. To curtail class imbalance in the other three staff segments, we undersample the true class distribution: we only load staff segments that contain a rare symbol (which includes symbols generally found at the beginning of a staff) . For example, an image is discarded if the image contains only *(semi)minims*.
In order to distinguish major and minor prolation in mensuration signs of fusa's from semifusa's in our dataset, we can not use a neural net as a consequence of the relative rarity of these symbols. We map semifusas to the *fusa* super class and major prolation to the *minor prolation* super class. We train a random forest classifier to discern the classes within each of the aforementioned super classes. We use Hu moments as discriminating features. In a first pass model, we simplify the ligature annotations in the ground truth set and detect ligature notes using the *ligature note* super class label. A second pass is then performed on the region containing just the ligature super classes, this time with detailed ligature labels, containing staff color and staff orientation e.g. *lnote tail left up*, *lnote tail right down* so that we can parse them according to Apels heuristic. We find a separate model honing in on the ligatures to be necessary given the sub-par regression of bounding boxes of densely written ligature symbols (discouraging us from the use of random forests) and the low confidence of a non-specialized model subject to the Non Maximum Suppression algorithm. By grid searching on the validation dataset, we find that an overlap threshold of 45% and confidence threshold of 20% strikes the optimal balance between eliminating unwanted symbols and removing densely written notes. We do not consider detections of the class *dot* in the NMS algorithm, as these small symbols have a low confidence and their bounding boxes overlap with that of the symbol to its left. Therefore, dots are filtered among themselves with NMS and later added to the rest of the inferred data. Finally, we deploy an extra heuristic: if a group of symbols appears before a clef or after a custos and the mean confidence is lower than 30 % than the symbols are removed.

4.3 Staff margin detection and local staff detection

Using the output of the object detection neural network – i.e. bounding boxes and their corresponding object labels – we estimate the location and size of the staffs on the page.

4.3.1 Stave line and inter staff line spacing determination

We start from the bounding box which should encapsulate the whole staff. As a countermeasure to staffs not containing every stave line, we expand the upper staff margin to fully contain each bounding box of the symbols found on the staff. We crudely approximate the height of the inter staff line spacing by using the bounding box heights as a prior. Symbols are annotated such that elements in $stems = \{$(semi)minim, (semi)fusa $\}$ have a bounding box height equal to the staff height, while symbol in $nostems = \{$(colored)(semi)breve$\}$ have a box height equal to 2 staff spacings. The estimate is then given by

$$\text{est} = \sum_i \frac{w_i \cdot h_i}{\#stems + \#nostems,} \tag{1}$$

where h_i is the bounding box height of symbol i, $w_i = 1/2$ for symbols in the *nostems* set and $w_i = 1/5$ for symbols in the *stems* set.
This crude estimate is used to obtain a rectangular structuring element. With this element we dilate and erode the image around the bounding boxes to probe the horizontal staff structure with the appropriate inter staff spacing. Subsequently, we binarize the staff segment by means of Gaussian adaptive thresholding. From the resulting image we can locate the peaks in the ensuing histogram corresponding to stave lines. Let Δ_i be the difference in the height coordinates of the peaks, then in order to discern true spacings, we subject Δ_i to:

$$\left| \frac{\text{est} - \Delta_i}{\Delta i} \right| < t + j, \tag{2}$$

with t the thickness as a fraction of the inter staff spacing and j the jitter as a fraction of the inter staff spacing . We set the bounding hyper parameter $b := 0.3$ with $b = t + j$. The inter staff spacing is then given by the median of the sampled and valid spacings.

4.3.2 Estimating the middle of the staff

Staffs can be skewed and/or degraded. Staff skewness prevents us from using a global estimate of the staff: the larger the staff gradient the smaller the window is where a constant middle estimate is applicable. On the other hand, degradation compels us to have an adequately large sample of the staff in order to find enough stave line evidence.
Ink fading is rarely uniform, therefore, we sample the staff in three windows, among which we choose the one with the most confident middle estimate. These windows are excerpts from the beginning, middle and end of the staff. Each of these windows is cut in vertical slices or samples, the width w_{sample} of the box from which we extract stave lines is a hyper parameter set to a third of the inter staff spacing. The height of a sample in a window spans the height of the staff. The second hyper parameter for the model is the amount of samples of the window and thus the width of the window: through grid searching we obtain a value twelve samples, given the slicing frequency of

three samples for every distance of one staff spacing travelled along the staff width. For each of the twelve samples per window we determine the horizontal lines with the algorithm described in the previous section.

Every horizontal line i in a sample j induces a staff i.e. a group of five lines, the extra lines can be above and or below the middle candidate. The middle line candidate $\textit{middle}_{ij}$ gets a score s_{ij} based on its reconstructed staff. We require every induced line k_{ij} in the group around the middle to be fully contained in the previously determined staff box upper and lower margins, if not $s_{ij} = 0$. Else the score of a middle line candidate is given by

$$s_{ij} = \sum_{k_{ij}} \sigma(k_{ij}) \tag{3}$$

with

$$\sigma(k_{ij}) = \begin{cases} 1 & \text{if line } \mathrm{k}_{ij} \text{ is found and not faded} \\ 0.5 & \text{if line } \mathrm{k}_{ij} \text{ is faded} \\ 0 & \text{if line } \mathrm{k}_{ij} \text{ is not found} \end{cases} \tag{4}$$

Where we count a line as found if it is present as peak in the histogram. A faded horizontal line defined to be a line which is more faint than half the mean grayscale line intensity found in the box. The score s_i of a line i is then given by $s_i = \sum_j s_{ij}$. Middles within bound b are identified with one another and their scores summed. The middle is then given by highest scoring line among the three windows.

4.3.3 Full staff reconstruction

The staff is reconstructed by propagating the middle along the staff width and updating the positions of the constituent lines in accordance with the stave line evidence in each sample. The lines are updated according to

$$l_{m+1} = \frac{1}{a} \sum_{n=1}^{a} l_{m+n}, \tag{5}$$

with $|l_m - l_{m+1+n}| < b$ and hyper parameter $a = 5$, corresponding to the amount of samples we look ahead in determining l_{m+1}. This parameter was found through grid searching. In case of severe degradation where no lines are found the line position is maintained.

4.4 Pitch detection

Position detection of the other music symbols such as notes, clefs and the separation of the superclass of *rest* symbols into its subclasses is tasked to a small convolutional neural network, see Table 1. For each such symbol, the width of the image entered in the neural network is the rescaled width of the box, the image height is extended contain the whole staff along with half an inter staff spacing above and below as margin. We rescale the dimension of the resulting image to 100×100 pixels. To bypass erroneous labels as a consequence of emblems or text, we set every pixel above the original bounding box to white. As staffs can be degraded, we normalize the images further by explicitly drawing the inferred staff lines before feeding it to the neural network.

Layer type	Filters	Stride	Activation
Convolutional	$100 \times 3 \times 3$	1	ReLU
Convolutional	$64 \times 3 \times 3$	1	ReLU
Max pooling	2×2	2	
Convolutional	$128 \times 3 \times 3$	1	ReLU
Convolutional	$128 \times 3 \times 3$	1	ReLU
Max pooling	2×2	2	
Convolutional	$256 \times 3 \times 3$	1	ReLU
Convolutional	$256 \times 3 \times 3$	1	ReLU
Max pooling	2×2	2	
Linear	1024		ReLU
Linear	512		ReLU
Linear	15		Softmax

Table 1: The architecture of the neural network used for pitch detection. We use six convolutional layers activated with Rectified Linear Units activation functions, interleaved with three max-pooling layers to down-sample the spatial resolution, and three linear layers (also activated by ReLU functions) to produce a classification result.

	Recall
YOLOR	67.8%
Random forest	**93.8%**
KNN	85.2%

Table 2: The mean recall for parsing the mensuration sign and fusa/semifusa super classes with several methods. The YOLOR model serves as a baseline and is trained without the use super classes.

4.5 Musical interpretation

In mensural notation, the note duration is determined by the duration of adjacent notes, coloration, and dot of augmentation and division. We append the series of heuristics implemented by Thomae in order to convert the set of detected musical objects in mensural notation to Common Music Notation [3].

5. EVALUATION

We evaluate our model on 35 pages from the image dataset of 14 books and 2.932 pages. We obtain 50 pages, 18 pages a of which are from chansonniers, 10 are from choir books and 7 scans belong printed works. This test dataset contains 239 staffs with 6749 symbols spread across 35 different classes. Furthermore, we evaluate our work on the Symbolically Encoded Il Laurro Secc dataset (SEILS) containing 150 pages and 33 categories [22]. Of the latter dataset not a single page was seen during train or validation.

	No superclasses	Superclasses
Undersampling	97.22%	**97.41%**
No undersampling	97.19%	97.28%

Table 3: Comparison of the use of super classes and sub-sampling when detecting the duration of symbols. We observe optimal results when using both super classes and sub-sampling.

	pitch recall	duration recall	CER
Choirbooks	97.43%	98.45%	4.54%
Printed	97.54%	99.32%	4.21%
Chansonnier	95.18%	96.08%	8.89%
SEILS	97.25%	98.29%	4.56%

Table 4: Summary of the recall for the ensemble model with super classes and sub-sampling. Both pitch and duration recall are high. Unsurprisingly, best results on the printed works as they show a uniform handwriting style and generally display little signs of image degradation.

To evaluate the model we use the recall metric and the character error rate as given by Equation 6:

$$CER = 100 \cdot \frac{I + D + S}{N} \quad (6)$$

Where I is the number of insertions, D the number of deletions and S the number of substitutions made in producing a correct staff with N symbols.

In the test set, one staff was not detected, and of the detected staffs one was erroneously traced by our algorithm. Overall, the ensemble, i.e. the combination of the random forest and subsampled object detection models, produces a state-of-the-art average duration recall of 97.41% over the corpora displayed in Table 4. On the set of illuminated choir books, we achieve a duration recall of 97.43% on choir books as compared with Van Gool et al. who report a symbol extraction rate of 88.14%. On the SEILS dataset, we achieve comparable results to the mean 4.5% CER by Castellanos et al [23]. However, our model was not explicitly trained on the aforementioned dataset. Furthermore, we note from Table 3 how undersampling and super classes have a beneficial effect when dealing with class imbalance. From Table 2, we see how the model which does not use super classes on the rare symbols fares worse than both K-NN and random forest, the latter being the most performant by 8.6 % margin when disentangling the semifusa/fusa, mensuration symbol super classes. Unsurprisingly, we encounter the largest drop in model performance with chansonniers. For these sources, the character error rate is rises to 8.9 %. This is due to degradation and a less clear handwriting styles along with a denser spacing between symbols as compared to printed works. Furthermore, in our test set the specialized ligature model detects ligatures with 100 % accuracy; an example of the output of the object detection is seen in Figure 2.

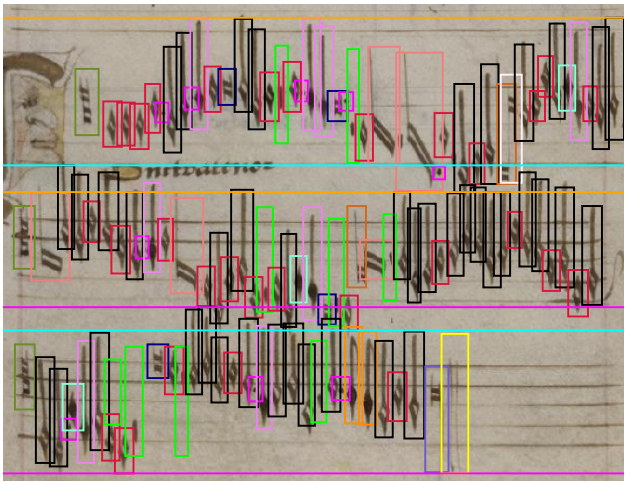

Figure 2: A crop of folium 10bis, from the Dijon Chansonnier [2]. We can see that – despite densely written notes and overlapping staff margins – our ensemble model correctly identifies all notes save for a single *rest*.

The YOLOR symbol detection model does not struggle with small symbols: the model detects the *dot* class with a recall of 97.8%, and *Rests* with a recall of 95.1%. The false negative *rests* are mostly due to non maximum suppression in densely written *rest* sequences.

Our model does produce some false positives. Most of these errors are due to overlapping symbols such as dots in clefs and the edge of staff being mistaken for a *custos*.

6. CONCLUSION

In this work, we demonstrated that inherent challenges of Optical Music Recognition on early documents such as image degradation, class imbalance, and small scale symbols can be successfully dealt with. Deviating from the classical OMR pipeline by Rebelo et al. [9], we used bounding box information to detect staff margins. Additionally, an algorithm was devised to deal with possibly degraded and skewed staffs. Finally, our model achieves state-of-the-art results on a variety of source types and handwriting styles. By incorporating the possibility for users to rectify mistakes and automatically transcribe works to Common Music Notation, we offer researchers in polyphonic music a way to process our collective knowledge of an important part of our cultural heritage. For the benefit of transparency and future OMR research, we make our annotated dataset publicly available.

In future work, we intend to incorporate a language model to incorporate corrections for the remaining fals positive symbols and erroneous pitch detections. Additionally, the method we propose can be adapted to detect plainchant notation, or modern notation alike. Applications of this system may include comparing the writing style of different scribes and digitizing music image databases.

7. ACKNOWLEDGMENTS

8. REFERENCES

[1] Alamire Foundation and Research Group Musicology KU Leuven, "Polyphony – IDEM Integrated Database for Early music," May 2021. [Online]. Available: https://idemdatabase.org/polyphony

[2] P. W. Christoffersen, "The Copenhagen Chansonnier and the 'Loire Valley' chansonniers An open access project," May 2021. [Online]. Available: http://chansonniers.pwch.dk/index.html

[3] M. E. Thomae, "Mensural-score-to-CMN-score," February 2020. [Online]. Available: https://github.com/martha-thomae/Mensural-score-to-CMN-score

[4] R. Timofte and L. V. Gool, "Automatic stave discovery for musical facsimiles," in *Computer Vision - ACCV 2012, 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part IV*, ser. Lecture Notes in Computer Science, K. M. Lee, Y. Matsushita, J. M. Rehg, and Z. Hu, Eds., vol. 7727. Springer, 2012, pp. 510–523. [Online]. Available: https://doi.org/10.1007/978-3-642-37447-0_39

[5] L. Pugin, "Editing renaissance music: The aruspix project," in *Digitale Edition zwischen Experiment und Standardisierung*, ser. Beihefte zu editio. Berlin, New York: Walter de Gruyter, 2009, vol. 31, pp. 147–156.

[6] P. Torras, A. Baró, L. Kang, and A. Fornés, "On the integration of language models into sequence to sequence architectures for handwritten music recognition," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021, Online, November 7-12, 2021*, J. H. Lee, A. Lerch, Z. Duan, J. Nam, P. Rao, P. van Kranenburg, and A. Srinivasamurthy, Eds., 2021, pp. 690–696. [Online]. Available: https://archives.ismir.net/ismir2021/paper/000086.pdf

[7] L. Pugin, J. Hockman, J. A. Burgoyne, and I. Fujinaga, "Gamera versus aruspix: Two optical music recognition approaches," in *ISMIR 2008, 9th International Conference on Music Information Retrieval, Drexel University, Philadelphia, PA, USA, September 14-18, 2008*, J. P. Bello, E. Chew, and D. Turnbull, Eds., 2008, pp. 419–424. [Online]. Available: http://ismir2008.ismir.net/papers/ISMIR2008_247.pdf

[8] Y. Huang, X. Chen, S. Beck, D. Burn, and L. Van Gool, "Automatic handwritten mensural notation interpreter: From manuscript to MIDI performance," in *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015, Málaga, Spain, October 26-30, 2015*, M. Müller and F. Wiering, Eds., 2015, pp. 79–85. [Online]. Available: http://ismir2015.uma.es/articles/191_Paper.pdf

[9] A. Rebelo, I. Fujinaga, F. Paszkiewicz, A. R. S. Marçal, C. Guedes, and J. S. Cardoso, "Optical music recognition: state-of-the-art and open issues," *Int. J. Multim. Inf. Retr.*, vol. 1, no. 3, pp. 173–190, 2012. [Online]. Available: https://doi.org/10.1007/s13735-012-0004-6

[10] J. Hajič jr., M. Dorfer, G. Widmer, and P. Pecina, "Towards full-pipeline handwritten OMR with musical symbol detection by u-nets," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018*, E. Gómez, X. Hu, E. Humphrey, and E. Benetos, Eds., 2018, pp. 225–232. [Online]. Available: http://ismir2018.ircam.fr/doc/pdfs/175_Paper.pdf

[11] F. J. Castellanos, J. Calvo-Zaragoza, G. Vigliensoni, and I. Fujinaga, "Document analysis of music score images with selectional auto-encoders," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018*, E. Gómez, X. Hu, E. Humphrey, and E. Benetos, Eds., 2018, pp. 256–263. [Online]. Available: http://ismir2018.ircam.fr/doc/pdfs/93_Paper.pdf

[12] J. A. Burgoyne, J. Devaney, L. Pugin, and I. Fujinaga, "Enhanced bleedthrough correction for early music documents with recto-verso registration," in *ISMIR 2008, 9th International Conference on Music Information Retrieval, Drexel University, Philadelphia, PA, USA, September 14-18, 2008*, J. P. Bello, E. Chew, and D. Turnbull, Eds., 2008, pp. 407–412. [Online]. Available: http://ismir2008.ismir.net/papers/ISMIR2008_221.pdf

[13] I. Barbancho, C. Segura, L. J. Tardón, and A. M. Barbancho, "Automatic selection of the region of interest in ancient scores," in *Melecon 2010-2010 15th IEEE Mediterranean Electrotechnical Conference*. IEEE, 2010, pp. 326–331.

[14] A. Pacha, J. Calvo-Zaragoza, and J. Hajič jr., "Learning notation graph construction for full-pipeline optical music recognition," in *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, The Netherlands, November 4-8, 2019*, A. Flexer, G. Peeters, J. Urbano, and A. Volk, Eds., 2019, pp. 75–82. [Online]. Available: http://archives.ismir.net/ismir2019/paper/000006.pdf

[15] International Music Score Library Project, "Ghirlanda Sacra (Simonetti, Leonardo)," May 2021. [Online]. Available: https://imslp.org/wiki/Ghirlanda_sacra_(Simonetti,_Leonardo)#IMSLP55989

[16] W. Apel, *The notation of polyphonic music, 900-1600*. Medieval Academy of Amer, 1961, no. 38.

[17] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "You only learn one representation: Unified network for multiple tasks," *arXiv preprint arXiv:2105.04206*, 2021.

[18] T. Lin, "Labelimg," 2015. [Online]. Available: https://github.com/tzutalin/labelImg

[19] J. Calvo-Zaragoza, I. Barbancho, L. J. Tardón, and A. M. Barbancho, "Avoiding staff removal stage in optical music recognition: application to scores written in white mensural notation," *Pattern Analysis and Applications*, vol. 18, no. 4, pp. 933–943, 2015.

[20] T. S. Nazaré, G. B. P. da Costa, W. A. Contato, and M. Ponti, "Deep convolutional neural networks and noisy images," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - 22nd Iberoamerican Congress, CIARP 2017, Valparaíso, Chile, November 7-10, 2017, Proceedings*, ser. Lecture Notes in Computer Science, M. Mendoza and S. A. Velastin, Eds., vol. 10657. Springer, 2017, pp. 416–424. [Online]. Available: https://doi.org/10.1007/978-3-319-75193-1_50

[21] MetaAI. (2022) sota real-time object detection on coco. [Online]. Available: https://paperswithcode.com/sota/real-time-object-detection-on-coco

[22] E. Parada-Cabaleiro, A. Batliner, A. Baird, and B. W. Schuller, "The SEILS dataset: Symbolically encoded scores in modern-early notation for computational musicology," in *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, October 23-27, 2017*, S. J. Cunningham, Z. Duan, X. Hu, and D. Turnbull, Eds., 2017, pp. 575–581. [Online]. Available: https://ismir2017.smcnus.org/wp-content/uploads/2017/10/14_Paper.pdf

[23] F. J. Castellanos, J. Calvo-Zaragoza, and J. M. Inesta, "A neural approach for full-page optical music recognition of mensural documents," in *Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR*, 2020, pp. 23–27.