

OPTICAL MUSIC RECOGNITION IN MENSURAL NOTATION WITH REGION-BASED CONVOLUTIONAL NEURAL NETWORKS

Alexander Pacha

Institute of Visual Computing and Human-Centered Technology, TU Wien, Austria
alexander.pacha@tuwien.ac.at

Jorge Calvo-Zaragoza

PRHLT Research Center
Universitat Politècnica de València, Spain
jcalvo@upv.es

ABSTRACT

In this work, we present an approach for the task of optical music recognition (OMR) using deep neural networks. Our intention is to simultaneously detect and categorize musical symbols in handwritten scores, written in mensural notation. We propose the use of region-based convolutional neural networks, which are trained in an end-to-end fashion for that purpose. Additionally, we make use of a convolutional neural network that predicts the relative position of a detected symbol within the staff, so that we cover the entire image-processing part of the OMR pipeline. This strategy is evaluated over a set of 60 ancient scores in mensural notation, with more than 15000 annotated symbols belonging to 32 different classes. The results reflect the feasibility and capability of this approach, with a weighted mean average precision of around 76% for symbol detection, and over 98% accuracy for predicting the position.

1. INTRODUCTION

The preservation of the musical heritage over the centuries makes it possible to study a certain artistic or cultural paradigm. Most of this heritage exists in written form and is stored in cathedrals or music libraries [10]. In addition to the possible issues related to the ownership of the sources, this storage protects the physical preservation of the sources over time, but also limits their accessibility. That is why efforts are being made to improve this situation through initiatives to digitize musical archives [17,21]. These digital copies can easily be distributed and studied without compromising their integrity.

Nevertheless, this digitalization, which indeed represents a progress with respect to the aforementioned situation, is not enough to exploit the actual potential of this heritage. To make the most out of it, the musical content itself must be transcribed into a structured format that can be processed by a computer [6]. In addition to indexing

the content and thereby enabling tasks such as content-based search, this could also facilitate large-scale data-driven musicological analysis in general [39].

Given that the transcription of sources is extremely time-consuming, it is desirable to resort to automatic systems. Optical music recognition (OMR) is a field of research that investigates how to build systems that decode music notation from images. Regardless of the approach used to achieve such objective, OMR systems vary significantly due to the differences amongst musical notations, document layouts, or printing mechanisms.

The work presented here deals with manuscripts written in mensural notation, specifically with sources from the 17th century, attributed to the Pan-Hispanic framework. Although this type of mensural notation is generally considered as an extension of the European mensural notation, the Pan-Hispanic situation of that time underwent a particular development that fostered the massive use of handwritten copies. Due to this circumstance, the need for developing successful OMR systems for handwritten notation becomes evident.



Figure 1. A sample page of ancient music, written in mensural notation.

We address the optical music recognition of scores written in mensural notation (see Figure 1) as an object detection and classification task. In this notation, the symbols are atomic units,¹ which can be detected and categorized independently. Although there are polyphonic composi-

¹ Except for beamed notes, in which the beam can be considered an atomic symbol itself.



tions from that era, each voice was placed on its own page, so we can consider the notation as monophonic on the graphical level. Assuming the aforementioned simplifications allows us to formulate OMR as an object detection task in music score images, followed by a classification stage that determines the vertical position of each detected object within a staff. If the clef and other alterations are known, the vertical position of a note encodes its pitch.

We propose using region-based convolutional neural networks, which represent the state of the art in computer vision for object detection, and demonstrate their capabilities of detecting and categorizing the musical symbols that appear in the image of a music score with a high precision. We believe that this work provides a solid foundation for the automatic encoding of scores into a machine-readable music format like Music Encoding Initiative (MEI) [38] or MusicXML [15]. At present, there are thousands of manuscripts of this type that remain to be digitized and transcribed. Although each manuscript may have its own particularities (such as the handwriting style or the layout organization), the approach developed in this work presents a common and extensible formulation to all of them.

2. RELATED WORK

Most of the proposed solutions to OMR have focused on a multi-stage approach [34]. This traditional workflow involves steps that have been addressed isolatedly, such as image binarization [4,47], staff and text segmentation [44], staff-line detection and removal [5, 11, 46], and symbol classification [3, 30, 33]. In other works, a full pipeline is proposed for a particular type of music score [31, 32, 43].

Recent works have shown that the image-processing pipeline can largely be replaced with machine-learning approaches, making use of deep learning techniques such as convolutional neural networks (CNNs) [1, 16, 29, 45]. CNNs denote a breakthrough in machine learning, especially when dealing with images. They have been applied with great success to many computer vision tasks, often reaching or even surpassing human performance [18, 22]. These neural networks are composed of a series of filters that operate locally (i.e. convolutions, pooling) and compute various representations of the input image. These filters form a hierarchy of layers, each of which represents a different level of abstraction [20]. The key is that these filters are not fixed but learnt from the raw data through a gradient descent optimization process [23], meaning that the network can learn to extract data-specific, high-level features.

Here, we formulate OMR for mensural notation as an object detection task in music score images. Object detection in images is one of the fundamental problems in computer vision, for which deep learning can provide excellent solutions. Traditionally, the task has been addressed by means of heuristic strategies based on the extraction of low-level, general-purpose features such as SIFT [28] or HOG [7]. Szegedy and colleagues [8, 42] redefined the use of CNNs for object detection for the first time. Instead

of classifying the image, the neural network predicted the bounding box of the object within the image. Around the same time, the ground-breaking work of Girshick et al. [14] definitely changed the traditional paradigm. In their work, a CNN was in charge of predicting whether each object of the vocabulary appeared in selected bottom-up regions of the image. This scheme has been referred to as region-based convolutional neural network (R-CNN). Afterwards, several extensions and variations have been proposed with the aim of improving both the quality of the detection and the efficiency of the process. Well-known examples include Fast R-CNN [13], Faster R-CNN [37], R-FCN [24], SSD [27] or YOLO [35, 36].

In this work, we use these region-based convolutional neural networks for OMR, which are trained for the direct detection and categorization of music symbols in a given music document. Thereby allowing for an elegant formulation of the task, since the training process only needs score images along with their corresponding set of symbols and the regions (bounding boxes) in which they appear.

3. AN OMR-PIPELINE FOR MENSURAL SCORES

Music scores written in mensural notation share many properties with scores written in modern notation: the sequence of tones and pauses is captured as notes and rests within a reference frame of five parallel lines, temporally ordered along the x-axis with the y-axis representing the pitch of notes. But unlike modern notation, mensural scores are notated monophonically with a smaller vocabulary of only around 30 different glyphs, reducing the overall complexity significantly and thus allowing for a simplified pipeline that consists of only three stages. A representative subset of the symbols that appear in the considered notation is depicted in Table 1.





















Group	Symbol			
	Semibrevis	Minima	Col. Minima	Semiminima
Note				
Rest	Longa	Brevis	Semibrevis	Semiminima
				
Clef	C Clef	G Clef	F Clef (I)	F Clef (II)
				
Time	Major	Minor	Common	Cut
				
Others	Flat	Sharp	Dot	Custos
				

Table 1. Subset of classes from mensural notation. The symbols are depicted without considering their pitch or vertical position on the staff.

3.1 Music Object Detection

The first stage takes as input an entire high-quality image that contains music symbols. The entire image is fed into

a deep convolutional neural network for object detection and yields the bounding boxes of all detected objects along with their most likely class (e.g., *g-clef*, *minima*, *flat*).

3.2 Position classification

After detecting the symbols and classifying them, the second stage performs position classification of each detected object to obtain the relative position with respect to the reference frame (staff) which is required to recover a notes pitch. For this process, we extract a local patch from the full image with the object of interest in the center and feed the image into another CNN, which outputs the vertical position, encoded as shown in Figure 2.

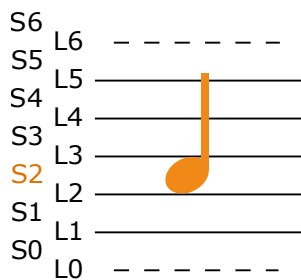


Figure 2. Encoding of the vertical staff line position into discrete categories. The five continuous lines in the middle form the regular staff and the dashed lines represent ledger lines, that are inserted locally as needed. A note between the second and third line from the bottom would be classified as S2 (orange).

3.3 Semantics Reconstruction and Encoding

Given the detected objects and their relative position to the staff line, the final step is to reconstruct the musical semantics and encode the output into the desired format (e.g., into modern notation [48]). This step has to translate the detected objects into an ordered sequence for further processing. Depending on the application and desired output, semantic rules need to be taken care of, such as grouping beams with their associated notes to infer the right duration or altering the pitch of notes when accidentals are encountered.

4. EXPERIMENTS

To evaluate the proposed approach, we conducted experiments² for the first two steps of the pipeline. While a full system would also require the third step, we refrain from implementing it, to not restrict this approach to a particular applications. It is also noteworthy, that translating mensural notation into modern notation can be seen as its own field of research that requires a deep understanding of

²Source code is available at <https://github.com/apacha/Mensural-Detector>

both notational languages, which exceeds the scope of this work.

4.1 Dataset

Our corpus consists of 60 fully-annotated pages in mensural notation from the 16th-18th century. The manuscript represents sacred music, composed for vocal interpretation.³ The compositions were written in music books by copyists of that time. To ensure the integrity of the physical sources, the images were taken with a camera instead of scanning the books in a flatbed scanner, leading to sub-optimal conditions in some cases. An overview of the considered corpus is given in Table 2.

Pages	60
Total number of symbols	15258
Different classes	32
Different positions within a staff	14
Average size of a symbol ($w \times h$)	44×84 pixels
Number of symbols per image	42–447 (\varnothing 250)
Image resolution ($w \times h$)	$\sim 3000 \times 2000$ pixels
Dots per inch (DPI)	300

Table 2. Statistics of the considered corpus.

The ground-truth data is collected using a framework, in which an electronic pen is used to trace the music symbols, similar to that of [2]. The bounding boxes of the symbols are then obtained by computing the rectangular extent of the users' strokes.

4.2 Setup

Our experiments are based on previous research by [29], where a sliding-window-approach is used to detect handwritten music symbols in sub-regions of a music score. In contrast to their work, we are able to detect hundreds of tiny objects in the full page within a single pass. To train a network in a reasonable amount of time within the constraints of modern hardware, it is currently necessary to shrink the input image to be no longer than 1000px on the longest edge, which corresponds to a downscaling operation by a factor of three on our dataset.

For detecting music objects, the Faster R-CNN approach [37] with the Inception-ResNet-v2 [41] feature extractor has been shown to yield very good results for detecting handwritten symbols [29]. It works by having a region-proposal stage for generating suggestions, where an

³The dataset is subject to ongoing musicological research and can not be made public at this point in time, so it is only available upon request.

object might be, followed by a classification stage, which confirms or discards these proposals. Both stages are implemented as CNNs and trained jointly on the provided dataset. The first stage scans the image linearly along a regular grid with user-defined box proposals in each cell of that grid.

To be able to generate meaningful proposals, the shape of these boxes has to be similar to the actual shape of the objects that should be found. Since the image contains a large number of very tiny objects (sometimes only a few pixels), a very fine grid is required. After a statistical analysis of the objects appearing in the given dataset, including dimension clustering [35], several experiments were conducted to study the effects of size, scale, and aspect ratios of the above-mentioned boxes, concluding that sensibly chosen priors for these boxes work similarly good as the boxes obtained from the statistical analysis. For the down-scaled image, boxes of 16x16 pixels, iterating with a stride of 8 pixels and using the scales 0.25, 0.5, 1.0, and 2.0, with aspect ratios of 0.5, 1.0, and 2.0 represent a meaningful default configuration. Accounting for the high density of objects, the maximum number of box proposals is set to 1200 with a maximum of 600 final detections per image.

For the second step of our proposed pipeline, another CNN is trained to infer the relative position of an object to its staff line upon which it is notated (see Figure 2). Different off-the-shelf network architectures are evaluated (VGG [40], ResNet [19], Inception-ResNet-v2 [41]) with the more complex models slightly outperforming the simpler ones. Using pre-trained weights instead of random initialization accelerates the training, improves the overall result, and is therefore used throughout all experiments. The input to the classification network is a 224 x 448 pixels patch of the original image with the target object in the center (see Figure 3). The exact dimensions of the patch are not important, as long as the image contains enough vertical and horizontal context to classify even symbols notated above or below the staff. When objects appear too close to the border, the image is padded with the reflection along the extended edge to simulate the continuation of the page as shown in Figures 3(d) and 3(e).

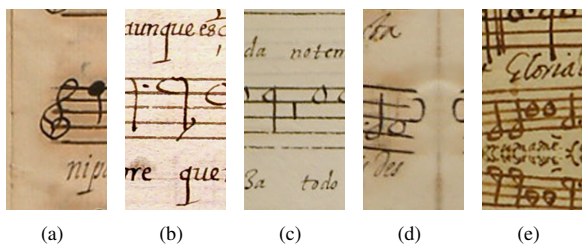


Figure 3. Sample inputs for the position classification network depicting a *g-clef* (a), *semiminima* (b), *brevis rest* (c), *custos* (d) and *semibrevis* (e), with vertical (d) and horizontal (e) reflections of the image to enforce the target object to be in the center, while preserving meaningful context.

It is important to notice that the vertical position defines the semantical meaning only for some symbols (e.g.,

the pitch of a *note* or the upcoming pitch with a *custos*). Classes for which the position is either undefined or not of importance include *barlines*, *fermatas*, different *time-signatures*, *beams* and in particular for mensural notation: the *augmentation dot*. Symbols from these classes can be excluded from the second step.

4.3 Evaluation metrics

Concerning the music object detection stage, the model provides a set of bounding box proposals, as well as the recognized class of the objects therein. The model also yields a *score* of its confidence for each proposal. A bounding box proposal B_p is considered positive if it overlaps with the ground-truth bounding box B_g exceeding 60%, according to the Intersection over Union (IoU) criterion: ⁴

$$\frac{\text{area}(B_p \cap B_g)}{\text{area}(B_p \cup B_g)}$$

If the recognized class matches the actual category of the object, it is considered a true positive, being otherwise a false positive. Additional detections of the same object are computed as false positives as well. Those objects for which the model makes no proposal are considered false negatives. Given that the prediction is associated with a score, different values of *precision* and *recall* can be obtained for each possible threshold. To obtain a single metric, Average Precision (AP) can be computed, which is defined as the area under this precision-recall curve. An AP value can be computed independently for each class, and then we provide the mean AP (mAP) as the mean across all classes. Since our problem is highly unbalanced with respect to the number of objects of each class, we also compute the weighted mAP (w-mAP), in which the mean value is weighted according to the frequency of each class. For the second part of the pipeline (position classification), we evaluate the performance with the accuracy rate (ratio of correctly classified samples).

5. RESULTS

Both experiments yielded very promising results while leaving some room for improvement. The detection of objects in the full image (see Figure 4) was evaluated by training on 48 randomly selected images and testing on the remaining 12 images with a 5-fold cross-validation. This task can be performed very well and yielded 66% mAP and 76% w-mAP. When considering practical applications, the weighted mean average precision indicates the effort needed to correct the detection results, because it reflects the fact that symbols from classes that appear frequently are generally detected better than rare symbols.

When reviewing the error cases, a few things can be observed: Very tiny objects such as the *dot*, *semibrevis rest* and *minima rest* pose a significant challenge to the network, due to their small size and extremely similar appearance (see Figure 5). This problem might be mitigated,

⁴ as defined for the PASCAL VOC challenge [9]



Figure 4. Detected objects in the full image with the detected class being encoded as the color of the box. This example achieves a mAP of approximately 68% and a w-mAP of 85%.

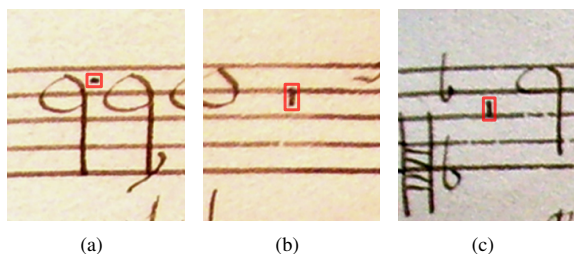


Figure 5. The smallest objects from the dataset that are hard to detect and often confused (from left to right): *dot*, *semibrevis rest*, and *minima rest*.

by allowing the network to access the full resolution image, which potentially has more discriminative information than the downsized image. Unsurprisingly, classes that are underrepresented such as *dots*, *barlines*, or all types of *rests* are also frequently missed or incorrectly classified, leading to average precision rates of only 10–40% for these classes.

Another interesting observation can be made, that in many cases, objects were detected but the IoU with the underlying ground-truth was too low for considering them a true positive detection (see Figure 6 with a red box being very close to a white box).

For the second experiment, a total of 13246 sym-

bols were split randomly into a training (80%), validation (10%) and test set (10%). The pre-trained Inception-ResNet-v2 model is then fine-tuned on this dataset and achieves over 98% accuracy on the test set of 1318 samples. Analyzing the few remaining errors reveals that the model makes virtually no errors and that the misclassified samples are mostly human annotation errors or data inconsistencies.

For inference, both networks can be connected in series. Running both detection and classification takes about 30 seconds per image when running on a GPU (GeForce 1080 Ti) and 210 seconds on a CPU.

6. CONCLUSION

In this work, we have shown that the optical music recognition of handwritten music scores in mensural notation, can be performed accurately and extendible by formulating it as an object detection problem, followed by a classification stage to recover the position of the notes within the staff. By using a machine learning approach with region-based convolutional neural networks, this problem can be solved by simply providing annotated data and training a suitable model on that dataset. However, we are aware that our proposal still has room for improvement. In future work we would like to:



Figure 6. Visualization of the performance of the object detection stage with selected patches of the music documents: green boxes indicate true positive detections; white boxes are false negatives, that the network missed during detection; red boxes are false positive detections, where the model reported an object, although there is no ground-truth; yellow boxes are also false positives, where the bounding-box is valid, but the assigned class was incorrect.

- evaluate the use of different network architectures, such as feature pyramid networks [25,26], that might improve the detection of small objects, which we have identified as the biggest source of error at the moment. These networks allow the use of high-resolution images directly, without the inherent information loss, that is caused by the downscaling operation.
- merge the staff position classification with the object detection network, by adding another output to the neural network, so the model simultaneously predicts the staff position, the bounding box and the class label.
- apply and evaluate the same techniques for other notations, including modern notation
- study models or strategies that reduce (or remove) the need for specific ground-truth data of each type of manuscript. For example, unsupervised training

schemes such as the one proposed in [12], which allows the network to adapt to a new domain by simply providing new, unannotated images.

We believe that this research avenue represents a ground-breaking work in the field of OMR, as the presented approach would potentially deal with any type of music scores by just providing undemanding ground-truth data to train the neural models.

7. ACKNOWLEDGEMENT

Jorge Calvo-Zaragoza thanks the support from the European Union’s H2020 grant READ (Ref. 674943), the Spanish Ministerio de Economía, Industria y Competitividad through Juan de la Cierva - Formación grant (Ref. FJCI-2016-27873), and the Social Sciences and Humanities Research Council of Canada.

8. REFERENCES

- [1] J. Calvo-Zaragoza and D. Rizo. End-to-End Neural Optical Music Recognition of Monophonic Scores. *Applied Sciences*, 8(4):606–629, 2018.
- [2] J. Calvo-Zaragoza, D. Rizo, and J. M. Iñesta. Two (note) heads are better than one: pen-based multimodal interaction with music scores. In *17th International Society for Music Information Retrieval Conference*, pages 509–514, 2016.
- [3] J. Calvo-Zaragoza, A. J. G. Sánchez, and A. Pertusa. Recognition of Handwritten Music Symbols with Convolutional Neural Codes. In *14th IAPR International Conference on Document Analysis and Recognition*, pages 691–696, 2017.
- [4] J. Calvo-Zaragoza, G. Vigiensoni, and I. Fujinaga. Pixel-wise binarization of musical documents with convolutional neural networks. In *15th IAPR International Conference on Machine Vision Applications*, pages 362–365, 2017.
- [5] J. S. Cardoso, A. Capela, A. Rebelo, C. Guedes, and J. P. da Costa. Staff detection with stable paths. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(6):1134–1139, 2009.
- [6] G. S. Choudhury, M. Droetboom, T. DiLauro, I. Fujinaga, and B. Harrington. Optical music recognition system within a large-scale digitization project. In *1st International Symposium on Music Information Retrieval*, pages 1–6, 2000.
- [7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005.
- [8] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. Scalable object detection using deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2147–2154, 2014.
- [9] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.
- [10] I. Fujinaga, A. Hankinson, and J. E. Cumming. Introduction to SIMSSA (Single Interface for Music Score Searching and Analysis). In *1st International Workshop on Digital Libraries for Musicology*, pages 1–3, 2014.
- [11] A.-J. Gallego and J. Calvo-Zaragoza. Staff-line removal with selectional auto-encoders. *Expert Systems with Applications*, 89:138–148, 2017.
- [12] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189, 2015.
- [13] R. Girshick. Fast R-CNN. In *IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.
- [14] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.
- [15] M. Good and G. Actor. Using MusicXML for file interchange. In *Third International Conference on WEB Delivering of Music*, page 153, 2003.
- [16] J. Hajič Jr. and P. Pecina. Detecting Noteheads in Handwritten Scores with ConvNets and Bounding Box Regression. *Computing Research Repository*, abs/1708.01806, 2017.
- [17] A. Hankinson, J. A. Burgoyne, G. Vigiensoni, A. Porter, J. Thompson, W. Liu, R. Chiu, and I. Fujinaga. Digital Document Image Retrieval Using Optical Music Recognition. In *Proceedings of the 13th International Society for Music Information Retrieval Conference*, pages 577–582, 2012.
- [18] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.
- [19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- [20] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, pages 1106–1114, 2012.
- [21] A. Laplante and I. Fujinaga. Digitizing musical scores: Challenges and opportunities for libraries. In *3rd International workshop on Digital Libraries for Musicology*, pages 45–48. ACM, 2016.
- [22] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
- [23] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [24] Y. Li, K. He, J. Sun, et al. R-FCN: Object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems*, pages 379–387, 2016.
- [25] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature Pyramid Networks for Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

- [26] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár. Focal Loss for Dense Object Detection. *Computing Research Repository*, abs/1708.02002, 2017.
- [27] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37, 2016.
- [28] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
- [29] A. Pacha, K.-Y. Choi, B. Coüasnon, Y. Ricquebourg, R. Zanibbi, and H. Eidenberger. Handwritten music object detection: Open issues and baseline results. In *13th IAPR Workshop on Document Analysis Systems*, pages 163–168, 2018.
- [30] A. Pacha and H. Eidenberger. Towards a Universal Music Symbol Classifier. In *12th IAPR International Workshop on Graphics Recognition*, pages 35–36, 2017.
- [31] L. Pugin. Optical music recognition of early typographic prints using hidden markov models. In *7th International Conference on Music Information Retrieval*, pages 53–56, 2006.
- [32] C. Ramirez and J. Ohya. Automatic recognition of square notation symbols in western plainchant manuscripts. *Journal of New Music Research*, 43(4):390–399, 2014.
- [33] A. Rebelo, A. Capela, and J. S. Cardoso. Optical recognition of music symbols. *International Journal on Document Analysis and Recognition*, 13(1):19–31, 2010.
- [34] A. Rebelo, I. Fujinaga, F. Paszkiewicz, A. R. Marcal, C. Guedes, and J. S. Cardoso. Optical music recognition: state-of-the-art and open issues. *International Journal of Multimedia Information Retrieval*, 1(3):173–190, 2012.
- [35] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.
- [36] J. Redmon and A. Farhadi. YOLO9000: Better, Faster, Stronger. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6517–6525, 2017.
- [37] 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*, pages 91–99, 2015.
- [38] P. Roland. The music encoding initiative (MEI). In *Proceedings of the First International Conference on Musical Applications Using XML*, pages 55–59, 2002.
- [39] X. Serra. The computational study of a musical culture through its digital traces. *Acta Musicologica*. 2017; 89 (1): 24-44., 2017.
- [40] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *Computing Research Repository*, abs/1409.1556, 2014.
- [41] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In *31st AAAI Conference on Artificial Intelligence*, pages 4278–4284, 2017.
- [42] C. Szegedy, A. Toshev, and D. Erhan. Deep Neural Networks for Object Detection. In *Advances in Neural Information Processing Systems 26*, pages 2553–2561, 2013.
- [43] L. J. Tardón, S. Sammartino, I. Barbancho, V. Gómez, and A. Oliver. Optical Music Recognition for Scores Written in White Mensural Notation. *EURASIP Journal on Image and Video Processing*, 2009(1):1–23, 2009.
- [44] R. Timofte and L. Van Gool. Automatic stave discovery for musical facsimiles. In *Asian Conference on Computer Vision*, pages 510–523, 2012.
- [45] E. van der Wel and K. Ullrich. Optical music recognition with convolutional sequence-to-sequence models. In *18th International Society for Music Information Retrieval Conference*, pages 731–737, 2017.
- [46] M. Visaniy, V. C. Kieu, A. Fornés, and N. Journet. The ICDAR 2013 music scores competition: Staff removal. In *International Conference on Document Analysis and Recognition*, pages 1407–1411, 2013.
- [47] Q. N. Vo, S. H. Kim, H. J. Yang, and G. Lee. An MRF model for binarization of music scores with complex background. *Pattern Recognition Letters*, 69:88–95, 2016.
- [48] Yu-Hui Huang, Xuanli Chen, Serafina Beck, David Burn, and Luc J. Van Gool. Automatic Handwritten Mensural Notation Interpreter: From Manuscript to MIDI Performance. In *16th International Society for Music Information Retrieval Conference*, pages 79–85, 2015.