ONE-STEP DETECTION OF BACKGROUND, STAFF LINES, AND SYMBOLS IN MEDIEVAL MUSIC MANUSCRIPTS WITH CONVOLUTIONAL NEURAL NETWORKS

Jorge Calvo-Zaragoza, Gabriel Vigliensoni, and Ichiro Fujinaga
Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT)
McGill University, Montreal, QC, Canada

ABSTRACT

One of the most complex stages of optical music recognition workflows is the detection and isolation of musical symbols. Traditionally, this goal is achieved by performing preprocesses of binarization and staff-line removal. However, these are commonly performed using heuristics that do not generalize widely when applied to different types of documents such as medieval scores. In this paper we propose an effective and generalizable approach to address this problem in one step. Our proposal classifies each pixel of the image among background, staff lines, and symbols using supervised learning techniques, namely convolutional neural networks. Experiments on a set of medieval music pages proved that the proposed approach is very accurate, achieving a performance upwards of 90% and outperforming common ensembles of binarization and staff-line removal algorithms.

1. INTRODUCTION

Optical music recognition (OMR) is the field of computer science devoted to providing computers with the ability to extract the musical content of a score from the optical scanning of its source image [1]. This problem represents a complex challenge for which there are no completely satisfactory solutions yet [20]. The task can be further divided into two different stages [6]: document image processing, in which the objective is to detect and recognize each meaningful symbol appearing in the image; and reconstruction of musical notation, in which musical meaning is assigned to each of these symbols in order to encode the content in a structured symbolic music format such as MEI (Music Encoding Initiative) or MusicXML.

Due to the arrangement of the elements on the staff, the image-processing stage is usually approached following a strategy of segmentation and classification. That is, elements within the score are first detected independently and then a classification algorithm assigns a category to each of them. Our approach focuses on the segmentation stage.

The final objective of segmentation is to detect the regions of the image that correspond to music symbols. To achieve this, traditional segmentation workflows incorporate the steps of binarization of the document and detection and removal of staff lines. Staff-line detection and removal algorithms usually use a binarized image as input, which facilitates certain procedures such as morphological operations or histogram analysis—core processes of many of these algorithms. In addition, the segmentation workflow also allows for the detection of staff line positions. If symbol isolation were done from a color image, it would not know which parts belong to the background of the document and which to staff lines. Note that the position of staff lines is crucial for determining the pitch of the symbols.

The traditional segmentation workflow, however, has a number of drawbacks. First, the staff-line detection and removal becomes heavily dependent on the accuracy of binarization, as errors are propagated between the two stages. In addition, the traditional methods follow heuristic techniques that assume specific conditions in the images to be treated. While this may be useful if the context of their use is limited to a particular style of documents, it is difficult to generalize these methods so that they can be used in various cases. This is especially true when dealing with medieval manuscripts, which present a greater heterogeneity in this regard.

For all of the above cases, we propose a framework with the goals of isolating the symbols depicted in the image of a music score and keeping the staff-line information. In our approach we perform a document analysis procedure that allows for categorical discrimination of each pixel, according to the class it belongs to (e.g., *background*, *staff lines*, or *symbols*) in a single step. In order to make this approach generalizable we address the task using the supervised learning paradigm. That is, we assume that a reference set is available that can be used to train a model to perform such task. In particular, we make use of convolutional networks for this purpose. These networks are powerful models that are capable of learning a suitable representation for a given task, thus avoiding the necessity of developing a feature extraction strategy specifically designed for each type of document to be processed. Our experiments on two sets of medieval documents report excellent results, outperforming different combinations of binarization algorithms and staff-line removal algorithms.

The rest of the paper is structured as follows: related work and the context of our proposal is presented in Section 2; the proposed method to solve the problem is detailed in Section 3; the experimental setup to validate our approach is described in Section 4; comparative and qualitative results are reported in Section 5; and conclusions and promising avenues for future work are summarized in Section 6.

2. RELATED WORK

OMR has to deal with many aspects of musical notation, one of which is the presence of the staff. Since most symbols in the score are connected through these lines, it has been traditionally necessary to remove them in order to detect musical symbols.

The staff-line removal stage is usually performed after the binarization of the document in the OMR workflow [20] because this step helps to reduce the complexity of the problem and is required to apply certain techniques such as morphological operators, histogram analysis, or connected components. In addition, starting from the color image, the processes of binarization and staff-line removal, one after the other, allow the separation of background, staff lines, and musical symbols regions.

A comprehensive review and comparison of the early attempts for the staff-line removal can be found in the work of Dalitz et al. [7]. Given the interest in this challenging task, many other methods have been proposed more recently. Cardoso et al. [9] proposed a method that considers the staff lines as connecting paths between the two margins of the score. The score was modeled as a graph so that staff detection was solved as a maximization problem. Dutta et al. [10] developed a method for printed scores that considered the staff-line segment as a horizontal connection of vertical black runs with uniform height. Piatkowska et al. [18] designed a method that used a Swarm Intelligence algorithm. Their approach can apparently deal with any type of image, but only results on binary images were reported. Su et al. [23] fitted an approximate staff considering properties such as height and space. Geraud [11] developed a method that entails a series of morphological operators: first, a permissive hit-or-miss with a horizontal line pattern, followed by a horizontal median filter and a dilation operation. A binary mask is then obtained with a morphological closing. Finally, a vertical median filter is applied to the largest components of the mask. The procedure is directly applied to the image, which eventually removes staff lines. Montagner et al. [15] proposed to learn image operators, whose combination remove staff lines.

The problem with these methods is that they focus on particular aspects of the style of the specific scores toward which they are oriented and it is, therefore, very difficult to adapt them to other types of documents (for example, from different eras or with different notations or styles). In addition, most of these methods assume already binarized images as input. The binary nature of modern musical scores (black ink on white paper) has, to some extent, justified this assumption. Of course, document binarization is not a trivial problem—especially when dealing with ancient documents [16]. Furthermore, it turns out that traditional document binarization methods, which were designed mainly for text documents, are often not suitable for musical scores [4].

Here we introduce a more generalized framework to solve the whole segmentation problem directly. The framework is based on machine learning so that it can be applied to a wide variety of musical notation styles and musical documents, as long as training data is available. Our strategy is inspired by the work of Calvo-Zaragoza et al. [5], in which a Support Vector Machine classifier was trained to discriminate if a *foreground* pixel of a binary image belongs to a *symbol* or to a *staff line*. Our approach is similar in formulation, but we do not assume that the documents are binarized or that they contain only symbols or staff lines. Furthermore, we also extend the procedure by using a more advanced classification scheme based on Convolutional Neural Networks (CNN).

3. METHOD

Although rarely formulated in this way, the problems related to image processing for musical documents are concerned with pixel-level classification processes. That is, for each pixel of the image we want to know whether it belongs to a musical symbol or not. In the latter case, we want to know whether the pixel belongs to a staff line or not, as this information is valuable for determining the vertical position of the notes (pitches), among others.

Therefore, the process can be formulated as a classification problem in which a model is trained to distinguish the category a given pixel belongs to. Formally, our approach considers a model that categorizes a given pixel into three possible classes: *background*, *staff*, and *symbol*. The requirement to carry out this idea consists of a reference set that allows providing examples of each category to the supervised learning algorithm.

In our framework, this classification process is carried out by means of Deep Learning. Recently, Deep Neural Networks have shown a remarkable leap of performance in pattern recognition. Specifically, CNN have been applied with great success for the detection, segmentation, and recognition of objects and regions in images, approaching human performance on some of these tasks [13].

CNN are composed of a series of filters (i.e., convolutions) that obtain several representations of the input image. These filters are applied in a hierarchy of layers, each of which represent different levels of abstraction; while filters of the first layers may enhance details of the image, filters of the last layers may detect high-level entities [12]. The key to this approach is that, instead of being fixed, these filters are modified through a gradient descent optimization algorithm called back-propagation [14].

One of the main advantages of CNN is their ability to learn a suitable representation of the training data without any human intervention, affording greater generalization to documents of different style. In other words, these networks learn a suitable representation of the data from raw data, without the need of feature extraction. Since collections of music documents are a rich source of highly complex information—often more heterogeneous than other types of documents—a framework based on CNN is promising.

3.1 Input Feature Set

As mentioned above, our intention is to train a CNN to differentiate pixels belonging to the different categories. Analyzing the organization of musical documents, we hypothesize that a pixel can be correctly categorized by using its local, neighboring information. In other words, we assume that the surrounding region of a pixel contains enough discriminative information to classify it into its correct category. As a result, the input set to the classifier in our framework is a portion of the input image, centered at the pixel of interest. Figure 1 illustrates some examples of input feature set for each of the considered categories, in which the pixel to be classified is located in the center of the patch.

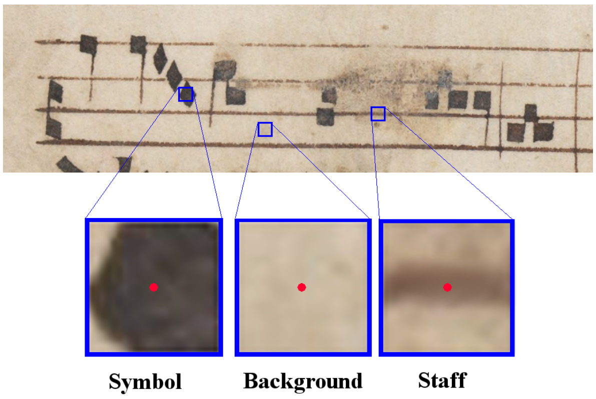


Figure 1. Example of input feature sets for pixels of interest of the three classes: *symbol*, *background*, and *staff*. Note that the pixel to be classified is located at the center of each window (highlighted in red for a better illustration).

Note that the method can work directly with color images and that the size of the neighborhood (i.e., the size of the window) is a parameter to be tuned according to the scale of the images to be processed.

3.2 Convolutional Neural Networks

Since there are no previously proposed CNN models to solve a task of this kind, we designed a new network configuration. Note, however, that the ultimate goal of this paper is not to find the best network topology—which would involve a comprehensive set of experiments to find the best set of parameters—but to demonstrate that the proposed categorization of music documents based on pixel-wise classification with CNN is feasible.

Our design is inspired by the VGG network [22], a topology widely used in the computer vision community for object recognition. This network contains several layers of convolution plus 2×2 max-pooling (16 or 19, depending on its version). By means of informal testing we simplified this network to up to 3 layers, adjusting the number of convolutional filters per layer to 64, and the size of the convolution kernels to 7.

Learning of the network weights is performed by means of stochastic gradient descent [2] with a batch size of 32, considering the adaptive learning rate proposed by Zeiler [26] (default parameterization) and a *cross-entropy* loss function. Once the CNN has learned how to distinguish among the considered categories it can be used to perform the layout analysis of a document. To do so, each pixel of the image is queried, and its feature set is forwarded and processed by the network in order to obtain its most likely category.

4. EXPERIMENTAL SETUP

4.1 Corpora

We trained and tested our approach on a set of high-resolution image scans of two different old music documents. The first corpus is a subset of 10 pages of the Einsiedeln, Stiftsbibliothek, Codex 611(89), from 1314. [1] The second corpus consists of 10 pages of the Salzinnes Antiphonal manuscript (CDM-Hsmu M2149.14), music score dated 1554–5. [2] Pages from the two manuscripts are shown in Figure 2. As a reference measure for scale, both pages depict a separation between staff lines of approximately 50 pixels.

Note that the image scans of these two manuscripts have zones with different lighting conditions that may affect the performance of the proposal we evaluate. The Einsiedeln manuscript images, in particular, present areas with severe bleed-through that may mislead the automatic recognition.

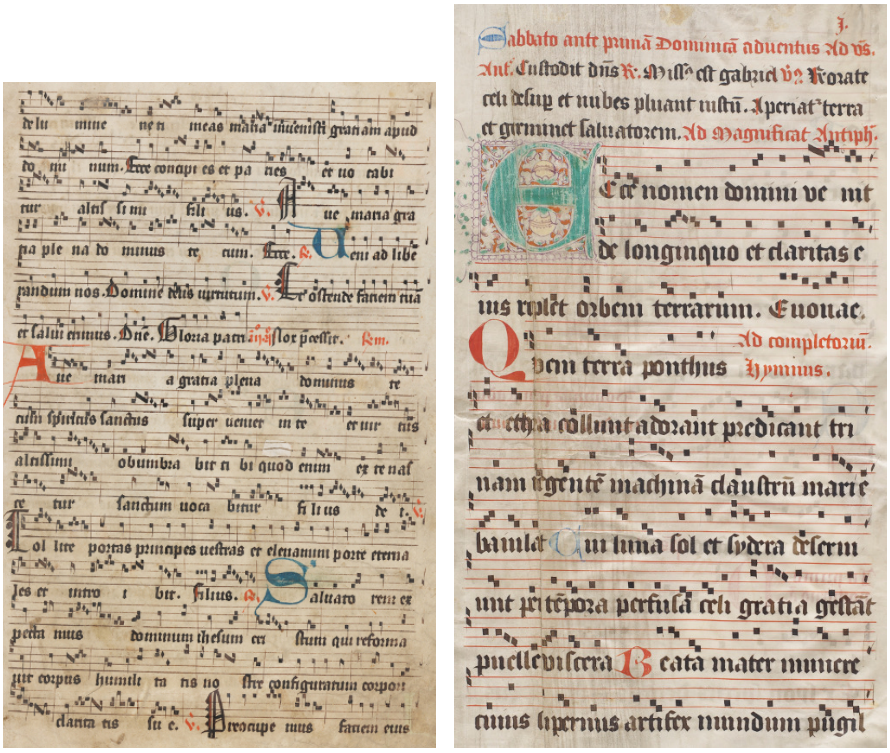
The ground-truth data from the corpora was created by manually labeling pixels into the three categories considered, as illustrated in Figure 3. Note that the class *symbol* includes both musical symbols and other types of symbols (such as lyrics). This should not be an issue as there exist successful algorithms to separate music and lyrics [3].

Taking into account the scale of the images of our corpora, an input window size of 41×41 pixels was empirically chosen, which corresponds to more than half of the space between the staff lines.

4.2 Comparative Assessment

To the best of our knowledge, there are no other algorithms that perform a direct detection of staff lines and symbols from music document images, and so we decided to compare our approach with combinations of standard binarization and staff-line removal algorithms. In order to select these algorithms, we took into account the results of the *ICDAR / GREC 2013 Competition on Music Scores: Staff Removal* [24]. In this contest, the two strategies that obtained the best performance were LRDE [11] and INESC [9].

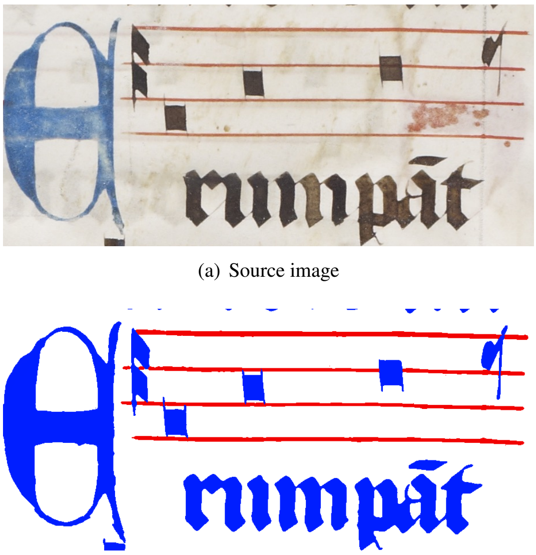
[1] http://www.e-codices.unifr.ch/en/sbe/0611/
[2] https://cantus.simssa.ca/manuscript/133/



(a) Einsiedeln

(b) Salzinnes

Figure 2. Pages from the corpora used in this work.



(a) Source image

(b) Ground-truth

Figure 3. Example of ground-truth created. Background pixels are labeled in white, staff-line pixels are labeled in red, and symbol pixels are labeled in blue.

These methods were based on published approaches (described in Section 2).

As mentioned before, these methods require that the input image only contains binary values. Therefore, the following binarization strategies are considered:

Sauvola method [21] is perhaps the most widely considered binarization algorithm for document images. It is based on the assumption that foreground pixels are closer to black than background pixels. It computes a threshold at each pixel considering the mean and standard deviation of a square window centered at the pixel under consideration.

Wolf & Jolion method [25] is an extension of Sauvola's, with a change in threshold formula to normalize contrast and the mean gray-level of the considered square window.

BLIST method [19] (Binarization based in LIne Spacing and Thickness) is specially designed for binarizing music scores. It consists of an adaptive local thresholding algorithm based on the estimation of the features of the staff lines depicted in the score.

To obtain the three categories mentioned above, we assume that *background* are those pixels removed by the binarization algorithm, while *staff* are those removed by the staff-line removal algorithm from the binarized image. The remaining pixels are thus classified as belonging to the *symbol* category.

Each combination of staff-line removal and binarization methods was evaluated experimentally. To assure a fair comparison, the parameters for each method (if any) were tuned to obtain the best results in the training set.

4.3 Evaluation

To evaluate our proposal, we used a scheme of 10-fold cross-validation on each corpus. That is, at each iteration, one of the pages was used as a test set, and the other nine were used to train the network and optimize its configuration. Specifically, 30 000 samples of each of the three classes were randomly selected for training (total: 90 000), while 600 000 of each class (total: 1 800 000) were used as a validation set. Note that these partitions represent a tiny portion of the available data, as each page contains about $2 \cdot 10^7$ pixels. However, these values were considered adequate to successfully train the networks (both in accuracy and computational load) on the machines that were used for that purpose. A more clever use of all available data will be discussed when addressing future work. As a result, the training set was used to optimize the CNN through gradient descent, whereas the validation set was used to select the most appropriate epoch to stop the learning process and prevent over-fitting. The complete testing pages were finally used to measure the performance of the model created by the network during training.

Given that the number of pixels of each class is not evenly balanced in the documents, we consider the F-measure (F_1) class-wise figure of merit for quantitatively assessing the classification accuracy of the system. Taking one class at a time as reference, this metric summarizes the correctly classified elements (*True Positive*, TP), elements falsely classified as belonging to the reference set (*False Positive*, FP), and elements of the reference set misclassified as belonging to another category (*False Negative*, FP) in a single value. Then, the F_1 is formalized as:

$$\mathrm{F}_1 = \frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \quad .$$

Finally, in order to minimize the possibility that the differences in model performance were due to chance variation, we will perform a pairwise, non-parametric test (Wilcoxon signed rank [8]).

5. RESULTS

We show in Table 1 the average F_1 results obtained in each corpus, as well as the overall performance when the whole set of documents is taken into consideration.

As can be seen in the table, the staff-line removal algorithm is the most relevant element in the considered configurations, because the differences are smaller when varying the binarization algorithm. In particular, the LRDE approach reports poor results in both sets of documents, despite having obtained the best results in the aforementioned competition. This directly demonstrates the lack of generalization of this approach. The INESC algorithm exhibits a fair performance, especially in the Salzinnes corpus. In regard to binarization algorithms, no conclusion can be drawn since the results seem too similar and depend on the corpus.

Strategy		Dataset		
		Einsiedeln	Salzinnes	Whole
LRDE	Sauvola	58.5	78.6	68.6
	Wolf	58.7	70.6	64.6
	BLIST	59.2	74.0	66.6
INESC	Sauvola	80.3	91.6	86.0
	Wolf	83.0	90.7	86.9
	BLIST	83.8	88.0	85.9
CNN		88.0	92.6	90.3

Table 1. Average F_1 obtained in the 10-fold cross-validation scheme for each corpus and the whole set.

The approach based on CNN, which performs the process in a single step, yields the best results in all cases considered. Since these results only reflect the average performance, we used the 20 independent results (10 for each corpus) to perform statistical tests. It resulted in p-values below 0.01 in all pairwise comparisons, and so our approach is significantly better than the rest of the configurations with an alpha significance level of 99%.

In order to have a qualitative reference, Table 2 shows an example of the categorization obtained by LRDE and INESC methods on a piece of Einsiedeln documents, considering BLIST binarization (best case), as well as the categorization of the approach based on CNN. It is observed that LRDE is only able to partially detect one of the lines of staff. INESC achieves an acceptable performance but it mislabels some sections of the staff. CNN shows a prediction that is very similar to the reference one. In addition, it completes one of the staff lines that is not perfectly seen in the original document (which, in turn, may be detrimental when computing its accuracy). Also, the CNN tends to mislabel pixel close to boundaries of elements, in which is not clear the actual category of the pixel. It is expected, however, that these errors will not cause inconveniences in subsequent procedures of the recognition workflow.

All in all, we can state that a trained CNN can successfully detect the selected categories at the pixel level in images of music scores. Our approach reports the best performance among the evaluated methods although it is fair to say that it is not by a wide margin. Nevertheless, its strength can be observed in the improvements achieved in each corpus. On the Salzinnes corpus, which seems to be less degraded and simpler, the margin was narrower. However, in the Einsiedeln manuscript the improvement over the compared methods was higher. This could mean that, as the difficulty increases, our approach could be more generalizable and adaptable.

It should be emphasized that the intention of this work was not to find the most suitable combination of input feature size and network topology, but to show that this approach allows dealing with the analysis of music doc-

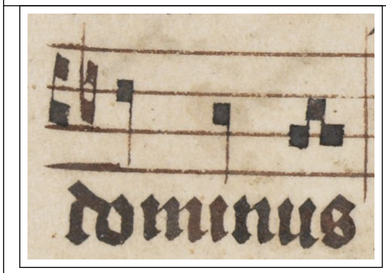
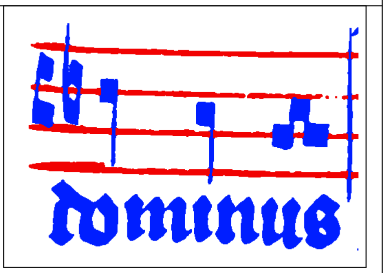
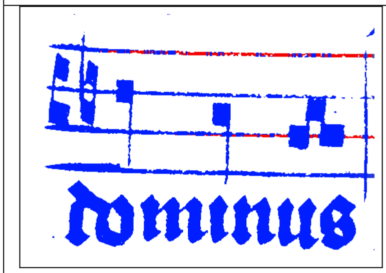
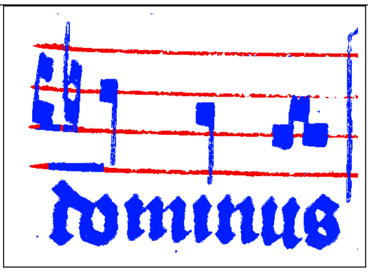
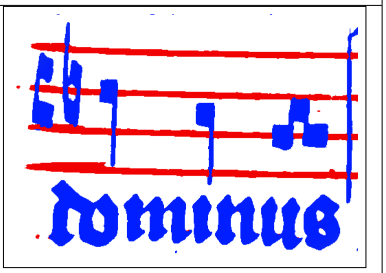
Source	Ground-truth	
		
BLIST+LRDE	BLIST+INESC	CNN
		

Table 2. Qualitative examples of categorization from Einsiedeln document, depicting the original piece along with the manually created ground-truth, and the labeling predicted by BLIST+LRDE, BLIST+INESC, and CNN. Coloring: background in white, staff lines in red, and symbols in blue.

uments successfully. Therefore, a more comprehensive search of the optimal parameters could be carried out to obtain an even better performance.

6. CONCLUSIONS

In this paper we presented a framework for detecting background, staff lines, and symbols in medieval manuscripts. Our approach was based on the classification of the different elements of the image at pixel level using machine learning. We use a CNN along with a training dataset of reasonable size that contained examples for each category.

Our results showed that the accuracy obtained is high, achieving around 90% of F_1 in the evaluated corpus. It has also been shown that our proposal is able to outperform state-of-the-art strategies based on heuristic image processing, demonstrating that CNN is a robust and generalizable alternative to those traditional approaches.

In future work, efforts should be devoted to overcoming the problem of getting enough data to train the CNN. It could be interesting to consider an incremental interactive framework in which the user does not have to label every single pixel of the image but only those erroneously labeled by a base classifier. The use of transfer learning [17] is another way to reduce the initial effort when dealing with a new type of document.

Moreover, there are several ways to improve the accuracy of the model in the future. Of course, finding a more suitable network configuration for this problem is a way of improving the results presented here. Also, since the available data is very large (i.e., a single page of ground-truth provides millions of examples of pixels labeled by humans), it would be more beneficial to train the network following a smarter strategy than choosing a random subset of the available data. For example, a random training set can be initially chosen to perform a first training iteration (as in the case of this work). After that, training documents can be evaluated so that the network is re-trained only with those pixels that would be misclassified by the current model. In this way, the network would pay special attention to the most difficult cases.

7. ACKNOWLEDGEMENT

This work was partially supported by the Social Sciences and Humanities Research Council of Canada.

Special thanks to Vi-An Tran for manually labeling the different layers in all the manuscripts used for this research.

8. REFERENCES

[1] D. Bainbridge and T. Bell. The challenge of optical music recognition. *Computers and the Humanities*, 35(2):95–121, 2001.

[2] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[3] J. A. Burgoyne, Y. Ouyang, T. Himmelman, J. Devaney, L. Pugin, and I. Fujinaga. Lyric extraction and recognition on digital images of early music sources. In *Proceedings of the 10th International Society for Music Information Retrieval Conference*, pages 723–727, 2009.

[4] J. A. Burgoyne, L. Pugin, G. Eustace, and I. Fujinaga. A comparative survey of image binarisation algorithms for optical recognition on degraded musical sources. In *Proceedings of the 8th International Conference on Music Information Retrieval*, pages 509–512, 2007.

[5] J. Calvo-Zaragoza, L. Micó, and J. Oncina. Music staff removal with supervised pixel classification. *International Journal on Document Analysis and Recognition*, 19(3):211–219, 2016.

[6] J. Calvo-Zaragoza, G. Vigliensoni, and I. Fujinaga. Document analysis for music scores via machine learning. In *Proceedings of the 3rd International Workshop on Digital Libraries for Musicology*, pages 37–40. ACM, 2016.

[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*, 30(5):753–766, 2008.

[8] J. Demsar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[9] J. Dos Santos 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*, 31(6):1134–1139, June 2009.

[10] A. Dutta, U. Pal, A. Fornés, and J. Lladós. An efficient staff removal approach from printed musical documents. In *20th International Conference on Pattern Recognition, ICPR 2010, Istanbul, Turkey, 23-26 August 2010*, pages 1965–1968, 2010.

[11] T. Géraud. A morphological method for music score staff removal. In *Proceedings of the 21st International Conference on Image Processing (ICIP)*, pages 2599–2603, Paris, France, 2014.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *26th Annual Conference on Neural Information Processing Systems*, pages 1106–1114, 2012.

[13] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[14] 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.

[15] I. d. S. Montagner, R. Hirata, and N. S. T. Hirata. A machine learning based method for staff removal. In *Proceedings of the 22nd International Conference on Pattern Recognition (ICPR)*, pages 3162–3127, 2014.

[16] K. Ntirogiannis, B. Gatos, and I. Pratikakis. ICFHR2014 competition on handwritten document image binarization (H-DIBCO 2014). In *14th International Conference on Frontiers in Handwriting Recognition*, pages 809–813, 2014.

[17] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[18] W. Piatkowska, L. Nowak, M. Pawlowski, and M. Ogorzalek. Stafflines pattern detection using the swarm intelligence algorithm. In L. Bolc, R. Tadeusiewicz, L. J. Chmielewski, and K. Wojciechowski, editors, *Computer Vision and Graphics*, volume 7594 of *Lecture Notes in Computer Science*, pages 557–564. Springer Berlin Heidelberg, 2012.

[19] T. Pinto, A. Rebelo, G. A. Giraldi, and J. S. Cardoso. Music score binarization based on domain knowledge. In *Proceedings of the 5th Iberian Conference on Pattern Recognition and Image Analysis*, pages 700–708, Las Palmas de Gran Canaria, Spain, 2011.

[20] 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. *International Journal of Multimedia Information Retrieval*, 1(3):173–190, 2012.

[21] J. Sauvola and M. Pietikäinen. Adaptive document image binarization. *Pattern Recognition*, 33(2):225–236, 2000.

[22] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *Computing Research Repository*, abs/1409.1556, 2014.

[23] B. Su, S. Lu, U. Pal, and C. L. Tan. An effective staff detection and removal technique for musical documents. In *Proceedings of the 10th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 160–164, 2012.

[24] M. Visaniy, V. Kieu, A. Fornés, and N. Journet. ICDAR/GREC 2013 music scores competition: Staff removal. In *Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1407–1411, 2013.

[25] C. Wolf, J.-M. Jolion, and F. Chassaing. Text localization, enhancement and binarization in multimedia documents. In *Proceedings of the International Conference on Pattern Recognition*, volume 2, pages 1037–1040, 2002.

[26] M. D. Zeiler. ADADELTA: An adaptive learning rate method. *Computing Research Repository*, abs/1212.5701, 2012.