A MACHINE LEARNING FRAMEWORK FOR THE CATEGORIZATION OF ELEMENTS IN IMAGES OF MUSICAL DOCUMENTS

Jorge Calvo-Zaragoza
Software and Computing Systems
University of Alicante
Alicante, Spain
jcalvo@dlsi.ua.es

Gabriel Vigliensoni, Ichiro Fujinaga
Schulich School of Music
McGill University
Montréal, Canada
{gabriel,ich}@music.mcgill.ca

ABSTRACT

Musical documents may contain heterogeneous information such as music symbols, text, staff lines, ornaments, annotations, and editorial data. Before any attempt at automatically recognizing the information on scores, it is usually necessary to detect and classify each constituent layer of information into different categories. The greatest obstacle of this classification process is the high heterogeneity among music collections, which makes it difficult to propose methods that can be generalizable to a broad range of sources. In this paper we propose a novel machine learning framework that focuses on extracting the different layers within musical documents by categorizing the image at pixel level. The main advantage of our approach is that it can be used regardless of the type of document provided, as long as training data is available. We illustrate some of the capabilities of the framework by showing examples of common tasks that are frequently performed on images of musical documents, such as binarization, staff-line removal, symbol isolation, and complete layout analysis. All these are tasks for which our approach has shown promising performance. We believe our framework will allow the development of generalizable and scalable automatic music recognition systems, thus facilitating the creation of large-scale browsable and searchable repositories of music documents.

1. INTRODUCTION

Optical Music Recognition (OMR) is the branch of artificial intelligence focused on automatically recognizing the content of a musical score from the optical scan of its source. In comparison to similar tasks such as text recognition, this process can be quite difficult given the complexity of music notation and the wealth of information contained in these documents. In addition to the musical notes that are usually overlaid on the staff lines, music scores may also contain several types of heterogeneous information such as alterations, lyrics, decorations, or bibliographic information about the piece. Therefore, before any attempt of automatic recognition, it is important to detect and classify these elements into their corresponding categories.

In addition to the tasks of symbol recognition and classification, there are other OMR preprocessing operations that are less well known. For example, a common first step in OMR workflows is binarization. This process consists in separating the background (i.e., the superfluous part of the image) from the foreground (i.e., the relevant content), and is usually considered the starting point for the subsequent OMR steps. A typical task that follows the binarization process is the detection and removal of staff lines. Although these lines are necessary for human readability and music interpretation, most OMR workflows are based on detecting and removing the staff lines before doing the classification of the remaining elements in the score.

OMR preprocessing is a complex step. In the past few years, many researchers have proposed OMR algorithms, workflows, and systems that deal with specific tasks on music documents, such as binarization [1], staff-lines detection [2], frontispiece delimitation [3], measure recognition [4], extraction of lyrics [5], and page border removal [6]. These approaches were all based on heuristical rules tailored to the music corpus at hand and achieved varying performance. Music documents have a high level of heterogeneity and exhibit many sources of variability, such as image degradation, bleed-through, different notation types, handwritten styles, or ink differences, among others. Therefore, if OMR systems are implemented by taking advantage of specific characteristics of the documents, different algorithms may be needed when working with sources of different type. As a result, the implementation of these systems will lack of generalizability and may be one of the factors hindering the progress of OMR technology.

In order to ameliorate this situation, we propose a generalized framework that allows detecting the different layers (i.e., background, staves, music symbols, lyrics, and so on) from the image of a music score, regardless of the specific characteristics of the source document. Extending the idea initially proposed by Calvo-Zaragoza et al. [7] for detecting and removing staff lines by using machine learning, we propose an approach in which each pixel of the image is labeled according to the type of content it depicts.

In contrast to strategies based on heuristic image processing, the main advantage of using machine learning rests in its generalizability. While the former focuses on particular aspects of the scores—being therefore very difficult to

adapt to other documents—techniques based on machine learning only need examples of the new type of documents to generate a different model. In some cases, it is even possible to reuse already trained models in documents of similar nature, but with a different type or style, by using Transfer Learning techniques [8].

Until a few years ago, the main disadvantage of using machine learning systems was that they did not achieve good results for image recognition tasks. However, since the rise of Deep Learning [9], Convolutional Neural Networks (CNN) have completely changed the scenario, outperforming traditional techniques in these tasks [10].

The rest of this paper is structured as follows: in Section 2 we detail the proposed unified framework and the rationale behind it. In Section 3 we show examples of tasks that can be successfully performed with the proposed framework. Finally, in Section 4 we summarize the core ideas of our method and gives some hints about future work.

2. DESCRIPTION OF THE FRAMEWORK

The framework we propose is based on the categorization of each pixel of interest within the input image with the label that illustrates to which information layer it belongs. To perform this task, we make use of the supervised learning paradigm [11]. That is, it is assumed that there will be enough representative examples of each type of information layer to be able to create a model to categorize new, unseen examples. Three elements are therefore essential for implementing this approach: (i) a feature set for each pixel, (ii) a classification algorithm, and (iii) training data.

2.1 Feature set

The feature set must characterize appropriately the pixel to be classified. We assume that the region of pixels around a specific pixel contains enough discriminating information to classify it with success. In other words, we hypothesize that a pixel can be correctly categorized by using the local information surrounding it. For example, whereas areas with staff lines may usually indicate zones where music notation is, areas without staff may indicate that other content, such as ornaments or lyrics, may be present. Text and decorations are similar in the local sense, but different ink type, color, or pen trace may have been used. Our approach exploits these local features to correctly distinguish the categories of the different elements within a musical document.

Figure 1 shows three examples of features sets for different pixels of an image. The pixel to be classified is located at the center of each window. Note that the size of the neighborhood (i.e., the size of the window) is a parameter to be tuned empirically, as the performance is highly related to this value [7].

Depending on the task, it might be advisable to increase the size of the window so that the features are discriminative enough. For example, with a small window it is possible that the feature set of a *text* sample would not be very different from those of *musical symbol*. However, increasing the size of the window too much may lead to an increase in the complexity of the problem, which could make the CNN not learn the task correctly. In addition, as the size of the feature set increases, a more computational time is needed.

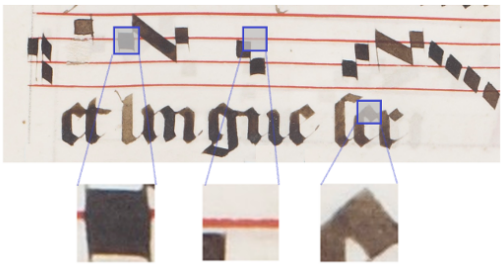


Figure 1. Example of feature sets from three regions of interest (i.e., music symbols, staff lines, and text). The pixel to be classified is located at the center of each window.

2.2 Classification algorithm

In our framework, the classification process is carried out by means of Deep Learning. Recently, Deep Neural Networks have shown a remarkable leap of performance in the field of machine learning. 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 [10].

These neural networks are composed of a series of filters (i.e., convolutions) that allow obtaining several representations of the input image. These filters are applied in a hierarchy of layers, each of which represent different levels of abstraction: whereas filters of the first layers enhance details of the image, filters of the last layers detect high-level entities [12]. The key is that these filters are not fixed but learned through a gradient descent optimization algorithm called back-propagation [13]. The configuration and organization of the network hierarchy (usually referred to as *topology*) has to be designed or chosen by the researcher.

Since collections of music documents are a rich source of highly heterogeneous information—usually more complex than other types of documents—developing a unified framework for OMR with a classification algorithm based on CNN is promising.

2.3 Training data

The last component to be considered in our framework is training data, which is dependent on the specific type of task to be performed. For example, it is likely that data needed to train a model to detect staff lines is different from data needed to discriminate among other items, such as musical symbols or text. Either way, the need of training data is the main drawback for the proposed framework, since it has to be created by manually labeling examples of all regions of interest in the document.

It is worth mentioning that we do not consider the possibility that a pixel belongs to more than one class at a time. We believe that from the point of view of an OMR system, in most cases there is just a single label that is truly rele-

vant. For example, pixels belonging to a musical symbol that are on a staff line should be considered as part of the former. But if needed, new categories for possible overlapping elements could be added, allowing the system to learn these categories as well.

3. EXAMPLES

In the following we present a number of examples of tasks in the classification of elements within musical documents. We made use of a CNN topology consisting of three convolutional layers. Although this might not be the best topology for the problems at hand, it is illustrative of the classification-based approach we propose. The window size of the feature set was specifically tuned for each example by means of informal testing.

The approach presented in this paper is directly applicable to any type of document no matter the type of notation and the style of the score, as long as enough training data is given to the network. In fact, different sources were considered for each of the examples in order to show how generalizable is our approach.

3.1 Binarization

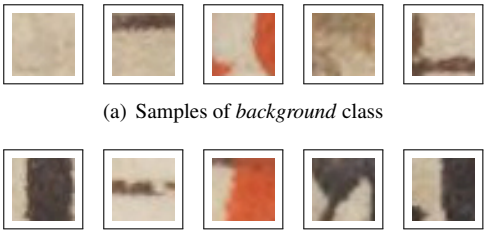
Binarization plays an important role in document analysis systems. This process is usually performed in the first stages of OMR systems and affects all subsequent stages. Therefore, it is crucial that binarization behaves in a robust way. Traditional binarization methods, however, have not shown consistent performance on music documents of different type. The degradation of music sources is one of the reasons for the unreliability of this process, but also great diversity in music notation is another obstacle [14].

The training data for this binarization example was composed of two manually labeled folios from Einsiedeln, Stiftsbibliothek, Codex 611(89). This manuscript is dated from 1314 and presents areas with severe bleed-through that may mislead standard binarization algorithms. From this labeled data, we selected the two layers of pixels that were labeled as *background* and *foreground*. We took random pixels from each layer and created a window of 25×25 pixels to be used as input feature for each pixel. We assumed that local information would be discriminative enough to classify correctly the center pixel. Figure 2 shows examples of features from both classes.

Once the CNN was trained with this data, it was able to distinguish between *background* and *foreground* pixels. As an illustrative example, Fig. 3 shows the binarization of a portion of a new document not seen during training that was classified pixel by pixel by the trained network. In spite of some spurious points that were misclassified, the network was able to achieve a remarkable performance for the binarization task.

3.2 Staff-lines detection and removal

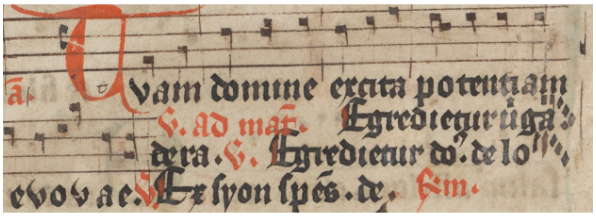
The detection and removal of staff lines follow the binarization step in most OMR workflows. Despite being necessary for musical readability, staff lines complicate the automatic detection, segmentation, and classification of symbols because they usually interconnect the symbols, thus not allowing their isolation.



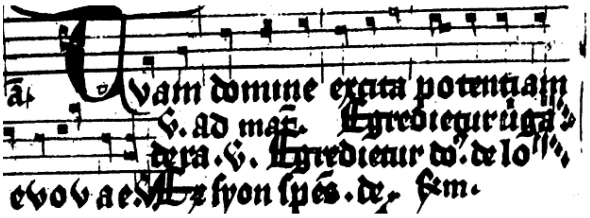
(a) Samples of *background* class

(b) Samples of *foreground* class

Figure 2. Training examples of both *background* and *foreground* classes. Each window has the pixel to be labeled at the center and also the local information to discriminate the class of the center pixel.



(a) Original input score portion



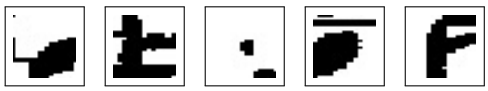
(b) Binarization of the input score

Figure 3. Example of binarization task performance achieved with our framework. The image was not part of the training set.

Traditional methods for the staff-lines removal task consider a binary image as input because it helps to reduce the complexity of the problem. In addition, binarization is mandatory for applying processes based on morphological operators, histogram analysis, or connected components. The binary nature of modern music scores (i.e., blank ink on white paper) have justified somewhat this workflow.

In this example, we show how the removal of staff lines from binary images can be performed successfully with our framework. We trained the network with a dataset that provided enough information to distinguish between pixels that belong to *staff* or *symbol* classes. In this case, we took advantage of the CVC-Muscima database [15] because it was a dataset especially designed for the evaluation of staff-lines removal tasks and contains handwritten common modern notation scores with and without staves. Figure 4 shows windows of pixels belonging to both classes.

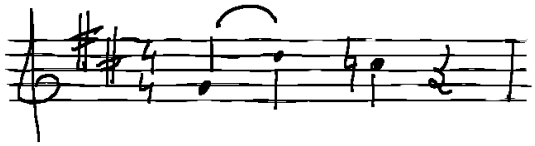
(a) Samples of *staff* class



(b) Samples of *symbol* class

Figure 4. Training data examples from *staff* and *symbol* classes.

We trained the CNN with enough data examples of the two classes, and then the network was able to detect and remove the staff lines accurately, as shown in Fig. 5.



(a) Example of input piece of score

(b) Input score after staff removal

Figure 5. Staff-lines removal task with binary images achieved with our framework.

3.3 Symbol isolation on color images

As introduced above, traditional methods for staff-line detection require a binary image as input. Since binarization processes are highly sensitive to conditions of the documents such as irregular lighting, image skewing, inkblots, or paper degradation, the performance of the symbol isolation task depends largely on the previous steps of binarization and staff-line removal. Fortunately, if we detect both background and staff-lines at the same time, the approach we propose in this paper enables complete symbol isolation in just one step. As a result, for this task there are three possible categories to tag a pixel: *background*, *staff*, or *symbol*. The latter included both music symbols and text characters.

In order to demonstrate the adaptability of our framework, we decided to try a new set of musical documents, and so we trained the network with pixel samples from three full pages of the Salzinnes Antiphonal (CDM-Hsmu M2149.l4) manuscript. Figure 6 shows examples of features for each category. A window size of 29×29 pixels was considered for this task.

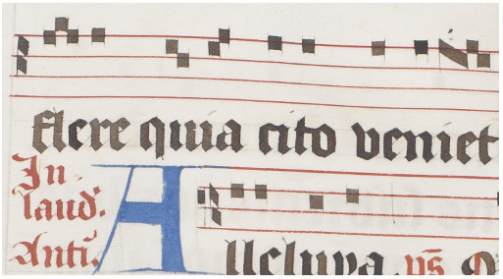
(a) Samples of *background* class

(b) Samples of *staff* class

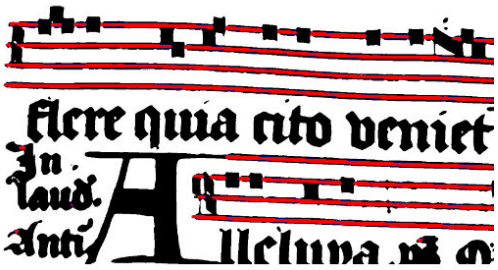
(c) Samples of *text* class

Figure 6. Examples of pixel windows from *background*, *staff*, and *symbol* classes.

After training the network, our approach was able to classify pixels belonging to the three different categories, as shown in Figure 7. The result is accurate but the detected staff lines are thicker than the original ones, possibly implying that the approach is over-sensitive in the local sense. The most plausible explanation is that the CNN does not notice too much difference amongst adjacent pixels, since the features are practically the same. This means that a pixel that is not on a staff line, but close to it, may be detected as a staff-line pixel by the network.



(a) Original input score portion



(b) Input score after staff-lines detection

Figure 7. Example of staff-lines detection on color images process achieved with our framework. Each layer considered is highlighted in a different color.

3.4 Music and text separation

Music symbols and text are important sources of information in music documents. Due to their different nature, text

and music are processed independently, with specialized automatic recognition algorithms. The proper separation of these two layers of information is a key aspect in the transcription of the whole document. We will show how our approach performs this classification task with ease.

In order to test the generalizability of our framework, we tested this task on a different music score, namely the GB-AR York Antiphonal manuscript. We manually classified pixels from one page into three different categories: *background*, *music*, and *text*. Fig. 8 shows a series of windows from each of these classes. Instead of a square window, preliminary experiments showed that a better perfomance was achieved with a rectangular window of 40×20 pixels.

(a) Samples of *background* class

(b) Samples of *music* class

(c) Samples of *text* class

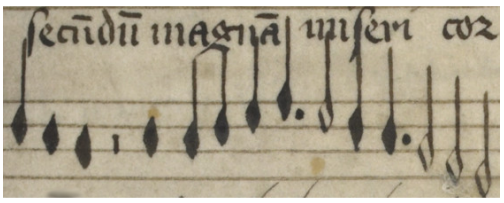
Figure 8. Examples of patches from *background*, *music*, and *text* classes.

Analogously to previous tasks, the CNN trained with these examples was able to produce accurate results, as shown in Figure 9. The framework achieved good performance even with those pixels where text and music symbols are overlapped. Nevertheless, as some pixels that belong to lyrics were erroneously classified as music, it is clear that the performance of this task is still not perfect.

This example highlights the strength of the framework we propose. It does not only separate text and music but it categorizes pixels at the pixel level—unlike previous approaches to this task that are devoted to just detecting zones or blocks of each type of information. Therefore, subsequent algorithms will not have to be in charge of performing the segmentation of the symbols within these blocks, since the specific pixels of interest are already detected.

3.5 Complete layout analysis

Typically, music scores contain much more information than just music symbols. This information includes titles, ornaments, lyrics, annotations, as well as unwanted artifacts such as ink bleed-through or ink blots. Therefore, a unified framework for complete document analysis of music documents should be able to identify and classify each of these categories within an input image. The framework we present in this paper is directly applicable to perform



(a) Example of input piece of score

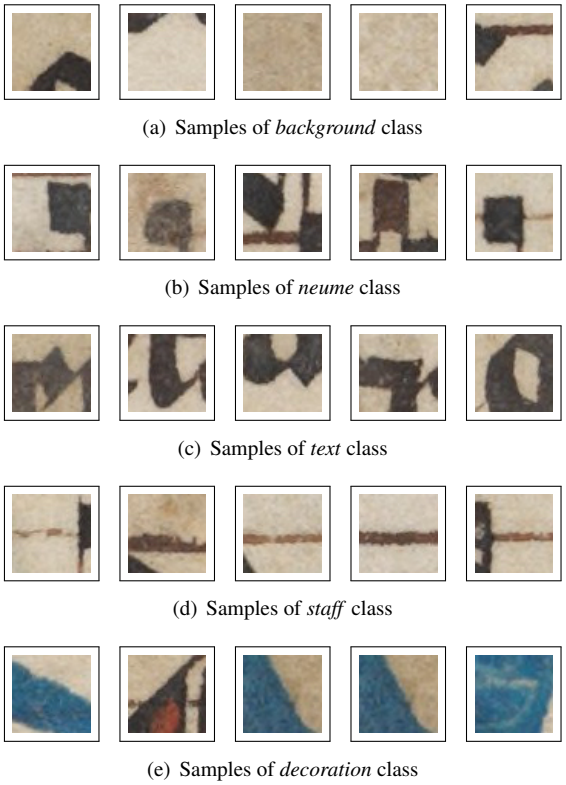
(b) Music and text separation in the input score

Figure 9. Detail of example of music and text separation using color images. Pixels classified as lyrics were labeled in red and symbols in black.

this task because it only needs enough training data and an appropriate window size surrounding each pixel.

Since the Einsiedeln manuscript contained several layers of interest within each page, such as music symbols, text, and ornamental letters, we tested a complete layout analysis in this manuscript. In this case, the data needed to be more discriminative and so we selected a window size of 51×51. As mentioned above, the specific size of the windows was chosen by performing preliminary experiments. What is important to remark in this case is that the window size needed to be larger than for the previous tasks because otherwise it would have been difficult to distinguish all categories. Also, since there were more categories, a larger amount of training data was required. Consequently, nine pages of the manuscript were manually labeled by categorizing their pixels into five different classes, namely *background*, *neume*, *text*, *staff*, and *decoration*. As a reference, the person in charge of building the training data required about 30 hours per page. Figure 10 shows a few examples of features extracted from this data, which were used to train the CNN.

Figure 11 shows an example of the categorization achieved by our framework. It can be seen that the result was not optimal, especially in the case of distinguishing between music symbols and text. Given the proximity of music and text, the feature windows for both categories were similar. Nevertheless, this example shows that a complete layout analysis is feasible, regardless of the categories to be considered, as long as training data is available and the feature window size is tuned accordingly. As mentioned at the beginning of this section, our intention was not to achieve the best classification results, but to determine how the framework may be applied in a different number of tasks and music documents. Further efforts on the parameterization of the classifier scheme (i.e., CNN topology, training data, and features) need to be carried out to achieve a better performance.



(a) Samples of *background* class

(b) Samples of *neume* class

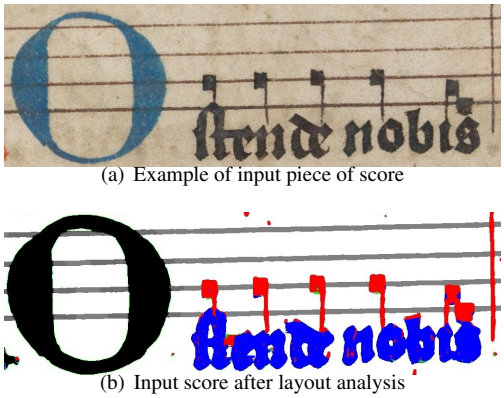
(c) Samples of *text* class

(d) Samples of *staff* class

(e) Samples of *decoration* class

Figure 10. Examples of window patches from all the categories considered for the complete layout analysis task.

Once all the different elements within the documents have been grouped into the corresponding categories, music symbols can be classified, text can be processed by Optical Character Recognition applications, and the positions of the staff lines and their corresponding clefs can be used to determine the pitch of notes. In addition, ornamental letters can be either removed to not disturb recognition algorithms or kept for extracting their meaning. As a side benefit, the background has been detected conveniently, helping to reduce the complexity of the recognition tasks.



(a) Example of input piece of score

(b) Input score after layout analysis

Figure 11. Detail of the complete layout analysis achieved by our framework on a previously unseen score. Each layer considered is highlighted in a different color.

4. CONCLUSIONS

In this paper we presented a unified framework for categorizing information contained in digitized images of music documents. Unlike previously proposed approaches for OMR tasks, our work presents a highly generalizable and scalable method that allows performing any task of image recognition in any kind of musical document.

Our system labels individual pixels of the image depending on the information they contain. To do so, the system uses machine learning techniques, namely CNN, to learn from examples of each category to be classified.

We showed different tasks that can be performed with our framework, such as document binarization, staff-lines removal in binary and color images, music symbols and text separation, and complete layout analysis. All these tasks can be solved directly by just changing the training data provided to the framework and tuning the window size considered as feature set.

We are aware that the categorization of every pixel and element in music documents is only a part of the whole OMR problem. However, we believe that the unified framework presented in this paper will allow the development of generalizable and scalable OMR systems, thereby enabling a breakthrough towards large-scale automatic recognition of heterogeneous music documents.

As future work, efforts should be devoted to overcoming the problem of getting enough data to train the CNN. For the examples showed above, training data was obtained manually. Since this may be too costly if needed for each new kind of document, a more efficient process must be pursued. For instance, labeled documents depicting different conditions—such as scale, deformations, and so on—could be generated synthetically in order to get representative examples of each type. The use of adaptive techniques for Domain Adaptation or Transfer Learning is another way to deal with this issue [16]. Furthermore, 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 [17].

Acknowledgments

This work was partially supported by the Social Sciences and Humanities Research Council of Canada and the Spanish Ministerio de Educación, Cultura y Deporte through a FPU Fellowship (Ref. AP2012–0939). Special thanks to Vi-An Tran for manually labeling the different layers in all the manuscripts used for this research.

5. REFERENCES

[1] 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*, 2011, pp. 700–8.

[2] C. Dalitz, M. Droettboom, B. Pranzas, and I. Fujinaga, "A comparative study of staff removal algorithms,"

IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 5, pp. 753–66, 2008.

[3] C. Segura, I. Barbancho, L. J. Tardón, and A. M. Barbancho, "Automatic search and delimitation of frontispieces in ancient scores," in *18th European Signal Processing Conference*, 2010, pp. 254–8.

[4] G. Vigliensoni, G. Burlet, and I. Fujinaga, "Optical measure recognition in common music notation," in *Proceedings of the 14th International Society for Music Information Retrieval Conference*, 2013, pp. 125–30.

[5] 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*, 2009, pp. 723–7.

[6] Y. Ouyang, J. A. Burgoyne, L. Pugin, and I. Fujinaga, "A robust border detection algorithm with application to medieval music manuscripts," in *Proceedings of the 2009 International Computer Music Conference*, 2009, pp. 101–4.

[7] J. Calvo-Zaragoza, L. Micó, and J. Oncina, "Music staff removal with supervised pixel classification," *International Journal on Document Analysis and Recognition*, vol. 19, no. 3, pp. 211–9, 2016.

[8] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in Neural Information Processing Systems*, 2014, pp. 3320–8.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1097–1105.

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–44, 2015.

[11] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, NY: John Wiley & Sons, 2001.

[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*, 2012, pp. 1106–1114.

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[14] 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 Society for Music Information Retrieval Conference*, 2007, pp. 509–12.

[15] A. Fornés, A. Dutta, A. Gordo, and J. Lladós, "CVC-Muscima: A ground truth of handwritten music score images for writer identification and staff removal," *International Journal on Document Analysis and Recognition*, vol. 15, no. 3, pp. 243–51, 2012.

[16] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: A survey of recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 53–69, 2015.

[17] L. Pugin, J. A. Burgoyne, and I. Fujinaga, "MAP adaptation to improve optical music recognition of early music documents using Hidden Markov Models," in *Proceedings of the 8th International Society for Music Information Retrieval Conference*, 2007, pp. 513–6.