Pixel-wise Binarization of Musical Documents with Convolutional Neural Networks

Jorge Calvo-Zaragoza
University of Alicante, Alicante, Spain
jcalvo@dlsi.ua.es

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

Abstract

Binarization is an important process in document analysis systems. However, so far there are no solutions that can be successfully applied to musical documents. In fact, it is possible that there might never be a certain process that allows a correct binarization for every type of musical document, since in addition to typical problems such as irregular lighting or source degradation, the heterogeneity of musical documents is very high. In this work we approach the binarization of musical documents by means of machine learning. Our approach is based on training a Convolutional Neural Network that classifies each pixel of the image as either background *or* foreground*. Our results demonstrate that the approach is competitive with other state-of-the-art algorithms. It also illustrates the advantage of being able to adapt to any type of score by simply modifying the training set.*

1 Introduction

Binarization plays a key role in systems for automatic analysis of documents [10]. This process is usually performed in the first stages of documents analysis, and serves as a basis for subsequent steps, hence it has to be robust in order to allow the full analysis workflow to be successful. When dealing with old music documents, the binarization step may be even more relevant.

The performance of most processes in Optical Music Recognition (i.e., the automatic transcription of musical documents into a structured digital format) workflows, is closely linked to the performance achieved by binarization. For example, a commonly performed stage in these processes is the staff-lines[1] removal, which facilitates the isolation of the different musical symbols of the document. Although some works have tried to solve this problem directly on color images with little success [14], most staff-line removal methods work with binary images because it helps to reduce the complexity of the problem, and a binary format is mandatory for applying processes based on morphological operators, histogram analysis, or connected components. Although staff-line removal is one important stage, it is not the only image-processing step that requires a proper binarization, such as removal of page borders [8], lyrics extraction [1], detection of measures [13], and delimitation of frontispieces [12].

It turns out that traditional document binarization methods, which were designed mainly for text documents, are not optimized for musical scores [2]. The specific reasons for this lack of generalizability are quite diverse, but they are mainly due to the large heterogeneity in music notation and style. Therefore, many of the assumptions made for text documents are not applicable in the case of music.

To alleviate this problem, we propose a method for binarizing musical documents using machine learning techniques. The main advantage of using automatic learning lies in its ability to generalize, in comparison to systems based on hand-crafted image processing strategies. While the latter focuses on singular aspects of the documents to be analyzed—being therefore very difficult to adapt to other types of documents of different epoch, notation, or style—techniques based on supervised machine learning only need labeled examples of new documents to generate a model adapted to the new environment.

Until a few years ago the main disadvantage of using machine learning systems was that they did not achieve good results for this type of tasks. However, Convolutional Neural Networks (CNN) changed this situation by outperforming traditional techniques in a wide range of image tasks [6]. These networks take advantage of local connections, shared weights, pooling, and many connected layers to learn a data representation and to successfully solve image processing tasks. Although traditional neural networks such as Multi-Layer Perceptron have been tested for binarization tasks [3, 5], to our knowledge CNNs has not been considered for the task at issue.

The rest of the paper is structured as follows: we provide a detailed description of the framework in Section 2. We present a series of experiments to validate our premise in Section 3. Finally, we conclude our work and present paths for future work in Section 4.

2 Document Binarization with Convolutional Neural Networks

Formally, the binarization task can be defined as a two-class classification task at pixel level. Our strategy basically consists in querying each pixel of the image to classify it as *foreground* or *background*. To do this, we use representative data of each pixel of interest and a CNN trained to distinguish between these two categories.

2.1 Network topology

The topology of these networks can be very varied. Our selected architecture can be described as follows: we denote by $conv(k, c, s)$ a convolutional layer with kernel size k, number of filters c, and stride s and Rectifier Linear Unit activation. Similarly, we denote by $maxpool(k, s)$ a max-pooling layer with kernel size k and stride s. $dropout(r)$ is a dropout procedure with a

[1] The set of horizontal parallel lines upon which notes are placed in written music notation

ratio of dropped units r; and $fc(c)$ is a fully connected layer with c outputs. Then, our network architecture can be defined sequentially as:

$$conv(3,3,32) \rightarrow maxpool(2,2) \rightarrow$$
$$conv(3,3,32) \rightarrow maxpool(2,2) \rightarrow$$
$$dropout(0.25) \rightarrow fc(128) \rightarrow$$
$$dropout(0.5) \rightarrow fc(2)$$

This topology was inspired firstly by LeNet-5 [7]. We then tested several modifications over the number of layers, number of filters, kernel sizes, type of activations and the addition of dropout units. Our final topology was finally selected according to the performance achieved during preliminary experiments.

2.2 Input data

The neural network to be trained must distinguish if a pixel from a musical document image belongs to the background or not. For that we assume that the region surrounding the pixel of interest contains enough information to discriminate between these two cases.

Hence, the input to the network will be a portion of the input image centered at the pixel of interest (see Fig. 1). It is clear that the window size of the surrounding region has a relevant impact on the classification and it has to be adjusted depending on the images to binarize.

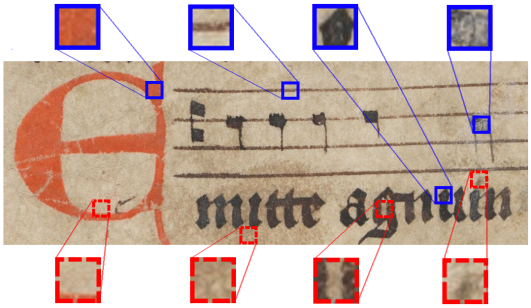


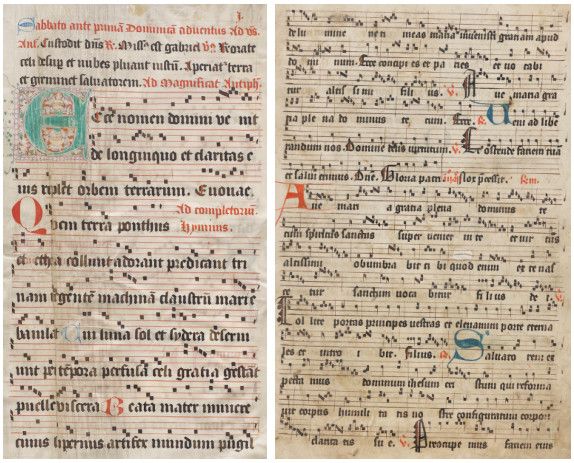
Figure 1: Examples of feature extraction from an old music document with a 17×17 square window for *foreground* (above, solid blue) and *background* (below, dashed red) categories. The pixel to be classified is located at the center of each window patch.

3 Experimental setup

In this section we detail the corpora we used, the metric for the evaluation, and we describe other binarization methods we used in our comparison with the same metric.

3.1 Corpora of musical documents

We trained and tested our approach on a set of high-resolution image scans of two different old music documents. The first corpus we tested was a subset of 10 pages of the Salzinnes Antiphonal manuscript (CDM-Hsmu M2149.14),[2] music score dated on 1554–5. The second corpus was 10 pages of the Einsiedeln, Stiftsbibliothek, Codex 611(89), from 1314.[3] Pages of these two manuscripts are shown in Fig. 2a and Fig. 2b, respectively. The image scans of these two manuscripts have zones with different lighting conditions, which may affect the binarization performance of the algorithms we evaluated. The Einsiedeln manuscript scans, in particular, present areas with severe bleed-through that may mislead standard binarization algorithms.



(a) Page from Salzinnes (b) Page from Einsiedeln

The ground-truth data from the corpora was created by manually labelling pixels into the two categories for the binarization task: *background* and *foreground*.

The size of the feature window was fixed to 17×17, which reported the best performance during preliminary experimentation.

3.2 Evaluation

In order to present reliable results, the experiments are carried out following a *leaving-one-out* cross-validation scheme at page level. That is, at each instance, one of the pages is left as test, whereas the training data comprise the rest of them.

For each fold, the size of the training set is fixed to 2 000 000 samples, randomly selected among the training pages. Note that this number of pixels only represents about 5 percent of the total number of pixels of any image of the corpora. Most of these samples (90 %) are used to optimize the CNN through gradient descent, whereas the rest is used as validation data to select the most appropriate epoch to stop the learning process and prevent over-fitting.

The complete testing page is used to measure the performance of the model created by the network during training. Since the number of foreground and background pixels is uneven, the performance metric considered is the F_1 score o(r *F-measure*).

3.3 Conventional binarization methods

To evaluate the benefit of our framework, we compared the results obtained by our approach with a few other algorithms widely used for binarization tasks:

Sauvola method [11] is based on the assumption that foreground pixels are closer to black than background

[2] https://cantus.simssa.ca/manuscript/133/

[3] http://www.e-codices.unifr.ch/en/sbe/0611/

pixels. It computes a threshold at each pixel considering the mean and standard deviation of a square window centered at the pixel at issue.

Wolf & Jolion method [15] is based on Sauvola, but it changes the threshold formula to normalize contrast and the mean gray-level of the piece of image.

Gatos method [4] is an adaptive procedure that follows several steps such as a low-pass Wiener filter, estimation of foreground and background regions, and a thresholding. It ultimately applies a post-processing step to improve the quality of foreground regions and preserve stroke connectivity.

BLIST method [9] (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 assure a fair comparison, the parameters of these methods (if any) were tuned by grid search using the same training set considered for the CNN.

3.4 Results

Average results obtained for each corpus are shown in Table 1.

Table 1: Average F_1 over a leaving-one-out validation for the corpora considered. *CNN* refers to the approach proposed in this paper. Values in bold represent the best results, on average.

Method	Dataset	
	Einsiedeln	Salzinnes
Sauvola	73.1	82.7
Wolf & Jolion	73.3	82.9
Gatos	60.3	78.9
BLIST	68.7	71.3
CNN	**77.1**	**84.3**

Analyzing the figures globally, we can see that the results obtained by our approach (CNN) were better than those obtained by the other methods in the two corpora. Since Wolf&Jolion and Sauvola methods are based on a similar scheme, they yielded similar results. The threshold tuning of the former may have improved the performance a bit. The Gatos method, traditionally reported as a good choice in text documents, shows a poor performance in these music documents. Finally, BLIST, the method specially designed for music scores, achieved poor results in both corpora. This method is tuned to the characteristics of modern music scores, and so it does not seem to be generalizable to the old music documents, as the ones we are considering. These results validate our initial premise that traditional binarization methods are not directly applicable to music documents. Also, it seems particularly beneficial to follow a machine learning approach in the case of musical scores, since the heterogeneity among different sources can be very high.

Qualitative results can be visualized in Fig. 3. It can be observed that our approach performed a binarization closer to the manually labeled ground-truth than any of the other methods. However, it tends to slightly dilate foreground regions, since pixels in the boundaries have similar features, and so the network is not able to distinguish them. Although this does not seem to cause visually errors, performance metrics are greatly degraded when compared to the ground-truth.

Although our approach is reporting the best performance, it is fair to say that they are not by a wide margin. Nevertheless, its goodness 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, with the Einsiedeln manuscript, the improvement over the other binarization methods was higher. This means that, as the difficulty increases, our approach may be more useful. Additionally, it should be emphasized that the intention of this work was not to find the most suitable combination of feature window sizes and network topology, but to show that this approach allows dealing with the binarization of musical documents successfully. Therefore, a more comprehensive search of the optimal parameters could be carried out to obtain even better results.

4 Conclusions

In this work we presented a new approach to binarize musical documents. Our strategy consist in training a CNN which then is capable of distinguishing background and foreground pixels.

Our experiments proved that our approach outperformed all the other widely used binarization methods. Further efforts on finding the best parameterization of the classifier scheme (i.e., the topology of the network, the training data, and features) might be be carried out to optimize the performance.

As future work our intention is to deal with the time-consuming problem of getting enough data to train the CNN. An interesting workflow to consider would be to create an initial training set by first using some existing heuristic binarization methods—such as those used above—which would then be manually edited to produce an appropriate ground-truth, hopefully more efficiently.

Acknowledgments

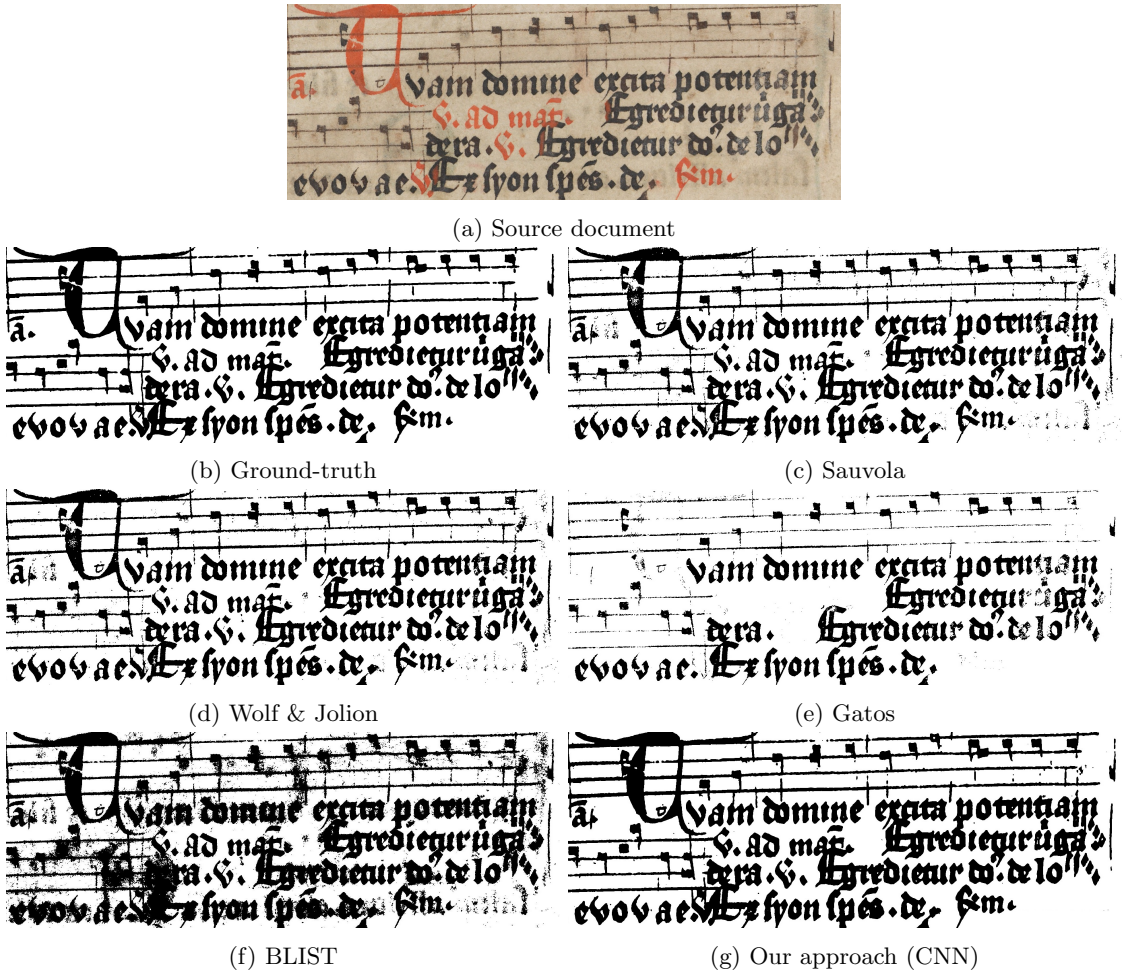
Temporary hidden due to anonymous revision.

References

[1] 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. *Proceedings of the 10th International Society for Music Information Retrieval*, pages 723–727, 2009.

[2] 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.

[3] Z. Chi and K. W. Wong. A two-stage binarization approach for document images. In *Proceedings of the International Symposium on Intelligent Multimedia, Video and Speech Processing*, pages 275–278. IEEE, 2001.



(a) Source document

(b) Ground-truth

(c) Sauvola

(d) Wolf & Jolion

(e) Gatos

(f) BLIST

(g) Our approach (CNN)

Figure 3: Qualitative comparison of the binarization methods considered on a piece of the Einsiedeln corpus.

[4] B. Gatos, I. Pratikakis, and S. J. Perantonis. Adaptive degraded document image binarization. *Pattern Recognition*, 39(3):317–327, 2006.

[5] A. Kefali, T. Sari, and H. Bahi. Foreground-background separation by feed forward neural networks in old manuscripts. *Informatica*, 38(4), 2014.

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

[7] 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.

[8] 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 International Computer Music Conference*, 2009.

[9] 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.

[10] I. Pratikakis, B. Gatos, and K. Ntirogiannis. ICDAR 2013 document image binarization contest (DIBCO 2013). In *Proceedings of the 12th International Conference on Document Analysis and Recognition*, pages 1471–1476, 2013.

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

[12] C. Segura, I. Barbancho, L. J. Tardón, and A. M. Barbancho. Automatic search and delimitation of frontispieces in ancient scores. In *Proceedings of the 18th European Signal Processing Conference*, pages 254–258, 2010.

[13] 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*, pages 125–130, 2013.

[14] M. Visaniy, V. C. Kieu, A. Fornes, and N. Journet. ICDAR 2013 Music Scores Competition: Staff Removal. In *Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1407–1411, 2013.

[15] 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.