Staff-line removal with Selectional Auto-Encoders

2 authors:

Antonio-Javier Gallego
University of Alicante
51 PUBLICATIONS   **275** CITATIONS

Jorge Calvo-Zaragoza
University of Alicante
80 PUBLICATIONS   **626** CITATIONS

Some of the authors of this publication are also working on these related projects:

Project   HispaMus: HISPAMUS: Handwritten Spanish Music Heritage Preservation by Automatic Transcription   View project

Project   MISIONES CRÍTICAS DE EMERGENCIAS CON MEDIOS AÉREOS TRIPULADOS Y NO TRIPULADOS EN VUELO COOPERATIVO (ENJAMBRE)   View project

Staff-line removal with Selectional Auto-Encoders

Antonio-Javier Gallego, Jorge Calvo-Zaragoza[*]

Department of Software and Computing Systems
University of Alicante, Carretera San Vicente del Raspeig s/n, 03690 Alicante, Spain

Abstract

Staff-line removal is an important preprocessing stage as regards most Optical Music Recognition systems. The common procedures employed to carry out this task involve image processing techniques. In contrast to these traditional methods, which are based on hand-engineered transformations, the problem can also be approached from a machine learning point of view if representative examples of the task are provided. We propose doing this through the use of a new approach involving auto-encoders, which select the appropriate features of an input feature set (Selectional Auto-Encoders). Within the context of the problem at hand, the model is trained to select those pixels of a given image that belong to a musical symbol, thus removing the lines of the staves. Our results show that the proposed technique is quite competitive and significantly outperforms the other state-of-art strategies considered, particularly when dealing with grayscale input images.

Keywords: Staff-line removal, Optical Music Recognition, Auto-encoders, Convolutional Networks

1. Introduction

Music is an important vehicle for cultural transmission, which is a key element as regards understanding the social, cultural and artistic trends of each

---

[*]Corresponding author: Tel.: +349-65-903772; Fax: +349-65-909326
Email addresses: `jgallego@dlsi.ua.es` (Antonio-Javier Gallego), `jcalvo@dlsi.ua.es` (Jorge Calvo-Zaragoza)

period of history. A large number of musical documents have, therefore, been carefully preserved over the centuries and are scattered throughout cathedrals, libraries or historical archives (Fujinaga et al., 2014).

A significant effort has been made in recent decades to digitize these documents by means of scanners for their storage and distribution (Choudhury et al., 2000). However, in order to make the music contained in the documents truly accessible, it is necessary for the images to be transcribed to a structured digital format that makes it possible to encode the content (notes, musical symbols, tonality, etc.) of the document. This also makes it possible to perform other interesting tasks, such as large-scale computational music analysis, search and retrieval by content, or transcription between different musical notations (Hankinson et al., 2012).

The process of converting a scanned document into a musical structured digital format can be carried out manually by a user. The disadvantage is that it involves costs in terms of both resources and time. In addition, this process is especially tedious—because of the burdensome software for score edition— and very prone to introducing errors.

The research field known as Optical Music Recognition (OMR), which focuses on detecting and storing the musical content of a score from a scanned image (Raphael & Wang, 2011), has therefore been postulated as an important alternative that mitigates the aforementioned disadvantages of manual transcription. The objective of the OMR process is to import a scanned musical score and export its musical content to a machine-readable format 1), typically MusicXML or MEI.

The OMR task is similar to the recognition of text, typically known as Optical Character Recognition. However, unlike the text scenario in which words are analyzed sequentially with a single element to identify at each time instant, musical notation is considerably more difficult to recognize. This is principally owing to the possible existence of simultaneous matching elements, as in the case of polyphonic pieces with multiple notes that sound at the same time, thus resulting in several musical symbols being placed on the same interval. But

2

(a) Example of input piece for an OMR system
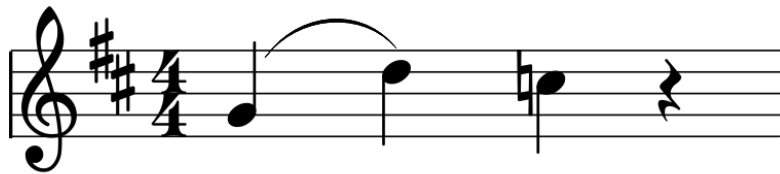
(b) Symbolic representation of the piece

Figure 1: The task of Optical Music Recognition (OMR) is to analyze an image containing a musical score in order to export its musical content to a machine-readable format.

it is also because of the presence of marks of expression, dynamics, articulations or even text to be sung in works with a vocal presence, among others.

Most current systems employ segmentation and classification approaches (Rebelo et al., 2010; Wen et al., 2015). The first important obstacle that the OMR process must overcome is, therefore, the staff (or pentagram) lines: the set of five parallel lines on which musical symbols are located depending on their pitch. The staff-line removal stage is usually performed after the binarization of the document in the OMR workflow (Rebelo & Cardoso, 2013). This binarization step helps to reduce the complexity of the problem, and it is advisable to apply strategies based on histogram analysis, connected components, or morphological operators.

Despite being necessary for musical readability, staff lines complicate the automatic detection and segmentation of symbols. Some specific works have taken advantage of specific features of printed and/or ancient notation to approach the

3

problem of maintaining the staff lines (Ramirez & Ohya, 2014; Calvo-Zaragoza

et al., 2015); however, the established OMR pipeline includes their detection and removal (Rebelo et al., 2012). This process must remove the staff lines while maintaining as much of the symbol information as possible (Fig. 2).

(a) Example of input score for an OMR system

(b) Input score after staff-line removal

Figure 2: Example of a perfect staff-line removal process.

This paper proposes a framework with which to remove staff lines that is based on machine learning, that is, labeled examples can be used to train a model to perform the task. This allows using the same approach in a wide range of scores. We make use of a new approach of auto-encoder, which is trained to select only those pieces of the image that belong to musical symbols.

The remainder of the paper is structured as follows: Section 2 presents the background to staff detection and removal; Section 3 describes our approach with which to model the process; Section 4 contains the experimentation performed and the results obtained; and finally, the current work concludes in Section 5.

4

2. Background

This section presents the background to our approach. First, methods proposed for staff-line removal are introduced, after which the principles of auto-encoders are briefly described, given their importance as regards the present work.

2.1. Staff-line detection and removal

Staff-line removal has been an active research field for many years. A comprehensive review and comparison of the first attempts considered for this task can be consulted in the work of Dalitz et al. (2008), who divided the staff-line removal strategies proposed until then into four categories: the Line Tracking (Randriamahefa et al., 1993; Bainbridge & Bell, 1997), Vector Field (Martin & Bellissant, 1991; Roach & Tatem, 1988), Runlength (Carter & Bacon, 1992; Fujinaga, 2005) and Skeleton (Ng, 2001) methods. However, given the interest in the task, many other methods have been proposed more recently.

Dos Santos Cardoso et al. (2009) proposed a method that considers the staff lines to be connecting paths between the two margins of the score. The score is then modeled as a graph so that staff detection is solved as a maximization problem. This strategy was later improved and extended to be used on grayscale scores (Rebelo & Cardoso, 2013).

Dutta et al. (2010) developed a method that considers the staff line segment as a horizontal connection of vertical black runs with a uniform height. These segments are validated using neighboring properties before removing them.

Su et al. (2012) started by estimating the properties of the staves such as height and space. An attempt is then made to predict the direction of the lines and fit an approximate staff, which is subsequently adjusted.

Géraud (2014) 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

5

to the largest components of the mask. The procedure is directly applied to the image, which eventually removes staff lines.

Notwithstanding all the efforts made, the staff-removal stage is still inaccu-
rate and often produces noise—staff lines not completely removed. Although it is possible to use more aggressive methods that minimize this noise, they may cause the partial or total loss of some musical symbols. The trade-off between these two aspects, in addition to the accuracy of the techniques, has hitherto led to the inevitable production of errors during this stage.

Moreover, the differences among score style, sheet conditions and scanning processes have led researchers to develop *ad-hoc* method for staff-line detection and removal. This results in methods that are not robust when applied to differ-
ent types of document (from different eras, or with different notations or styles). In modern notation, a staff is composed by five black parallel lines over a white background. However, that is not always true in old music notation because (i) staff lines may appear with different ink colors, even closer to background color than to symbol color, (ii) handwritten staff lines are not totally straight, (iii) the thickness of the lines is irregular because of quill leakage, and (iv) the staff does may have less than five lines. In addition, lyrics in modern scores are always far enough from the staff, whereas there is much overlapping be-
tween music and lyrics in old notation, and so it could also hinder the staff-line removal process. Therefore, many of the assumptions for staff-line removal in modern music are not always fulfilled in different eras, thereby being extremely difficult to develop methods that are able to work on any kind of scores with an acceptable accuracy.

Here we introduce a new and more generalized framework based on ma-
chine learning that can be applied to a wide variety of musical notation styles and musical documents. The main advantage of using machine learning lies in its ability to be generalized when compared with hand-crafted systems. In this respect, a machine learning strategy for staff-line removal has recently been pro-
posed, which consists of training a classifier that discriminates between whether a given pixel belongs to a symbol or to a staff line (Calvo-Zaragoza et al., 2016).

6

The foreground pixels of the image are queried so that those classified as *staff* are removed. Nevertheless, it has two important disadvantages, the first being that this strategy does not improve the performance of other state-of-the-art algorithms for staff-line removal. The second is its high computational cost, which results from having to classify each pixel in the image.

The aim of this work is to alleviate the aforementioned issues by using specialized auto-encoders. The following section presents a brief introduction to auto-encoders and their related extensions in order to provide the reader with a better understanding of the proposed framework.

2.2. Auto-encoders

Auto-encoders consist of feed-forward neural networks for which the input and output must be exactly the same. The network typically consists of two stages that learn the functions f and g, which are called encoder and decoder functions, respectively.

Formally speaking, given an input x, the network must minimize the divergence $L(x, g(f(x)))$. The hidden layers of the encoder perform a mapping of the input —usually decreasing its dimension— until an intermediate representation is attained. The same input is then subsequently recovered by means of the hidden layers of the decoder function.

An auto-encoder might initially appear to be useless because it is trained to learn the identity function. Nevertheless, the encoder function f is typically forced to produce a representation with a lower dimensionality than the input. The encoder function therefore provides a meaningful compact representation of the input, which might be of great interest as regards feature learning or dimensionality reduction (Wang et al., 2014).

The idea of auto-encoders was proposed decades ago (Hinton & Zemel, 1994), and it has since been an active research field (Deng et al., 2010; Lauly et al., 2014). As auto-encoders are feed-forward networks, they can be trained by using conventional optimization algorithms such as gradient descent.

In some applications, it is assumed that an input $\hat{x}$ is received after passing

7

through a noisy process that corrupted the actual input x. In this case, a *denoising auto-encoder* might be taught to minimize the divergence $L(x, g(f(\hat{x})))$. The network, therefore, not only focuses on copying the input but also on removing the noise (Vincent et al., 2010; Bengio et al., 2013).

In the context of staff-line removal, we could have formulated the problem by assuming that these lines are the result of noise and a denoising auto-encoder could, therefore have been taught to remove them. However, in this paper we go one step further and propose a new type of auto-encoder especially designed for the problem at hand. In this case, we wish neither to learn the identity function nor an underlying error but rather a codification that maintains only those input features that are relevant. That is, the output of our auto-encoder is a codification with the same dimension as the input that indicates the original features that must be maintained. Note that, in the staff-line removal formulation, the relevant features will be those pixels of the image that depict symbols, while those containing staff lines must be discarded.

3. Staff-line removal with Selectional Auto-Encoders

As mentioned above, the traditional formulation for the staff-line removal task considers a binary image as input; that is, a binarization process is assumed to be performed before this step. The binary nature of modern musical scores (black ink on white paper) has, to some extent, justified this pipeline. However, it should be borne in mind that document binarization is not a trivial question — especially when dealing with ancient documents (Ntirogiannis et al., 2014).

Furthermore, the staff-line removal stage is actually a process that attempts to leave only the musical symbols in the image. From this point of view, our proposal focuses on selecting the pixels of the input image that correspond to musical symbols, regardless of the nature of the input (see Fig. 3), thus leading to a more generalizable approach that can be applied to binary, grayscale or color images.

This has been done by making use of a new type of auto-encoder. As

(a) Traditional approach for isolating musical symbols in score images

(b) Proposed approach for isolating musical symbols in score images
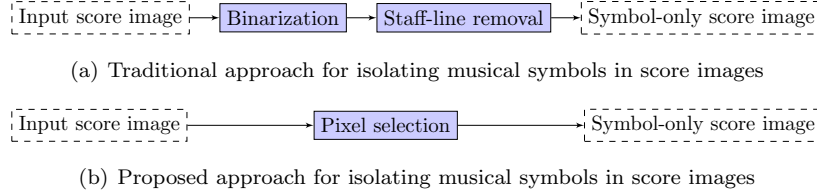
Figure 3: Comparison of traditional and proposed approaches for isolating symbols in musical score images.

mentioned above, the model is trained to select which features of the input layer are relevant for the task at hand (that is, the pixels that belong to music symbols). From here on, we shall refer to this model as the *Selectional* *Auto-Encoder* (SAE). The SAE is trained to perform a function such that $s : \mathbb{R}^{(w \times h)} \to [0,1]^{(w \times h)}$. In other words, it learns a binary map over a $w \times h$ image that preserves the input shape. Following the idea of auto-encoders, however, the function is further divided into encoding and decoding stages.

The topology of an SAE can be quite varied. However, we have restricted ourselves to considering convolutional models. Convolutional models have been applied with great success to the detection, segmentation and recognition of objects and regions in images, and have even come close to human performance in some of these tasks (LeCun et al., 2015). They can take advantage of local connections, shared weights, pooling and the use of many connected layers.

The hierarchy of layers of our SAE consists of a series of convolutional plus pooling layers, until an intermediate layer in which meaningful representations of the input are attained. As these layers are applied, filters are able to relate parts of the image that were initially far apart. It then follows a series of convolutional plus upsampling layers that reconstruct the image up to the same input size. The last layer consists of a set of neurons with sigmoid activation that predict a value in the range of $[0,1]$, depending on the *selectional* level for the corresponding input feature. The scheme of our hierarchy is illustrated in Fig. 4.

The specific configuration, along with a suitable size of the input/output

9

5x5x96
5x5x96
5x5x96

5x5x96
5x5x96
5x5x96

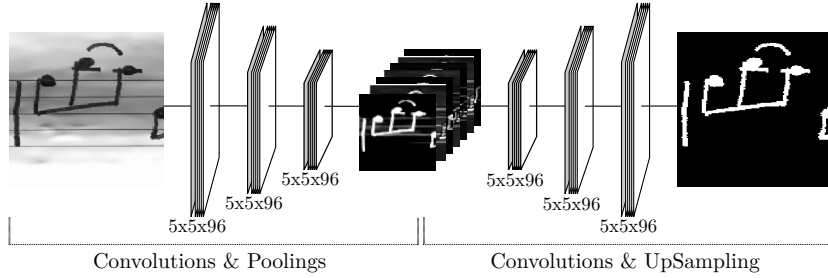Convolutions & Poolings   Convolutions & UpSampling

Figure 4: General overview of a SAE used for staff-line removal. The output layer consists of the activation level assigned to each input feature (white signifies activated)

layer, has been adjusted by means of a grid search of the network configuration hyper-parameters, as will be detailed in the next section. As a preliminary proof-of-concept test, we also carried out some research with non-convolutional models, which proved to be less suitable for the task at hand.

Since an SAE is a type of feed-forward network, the training stage is carried out by means of back-propagation, considering the *cross-entropy* loss function between each output activation and its expected activation. Let n be the number of training examples, and d be the dimensionality of the input (and output). Let us denote the activation of the jth neuron of the last layer for the example i as a_{ij} and its desired activation as y_{ij}. The loss L for the training set can therefore be computed as:

$$L = -\frac{1}{nd} \sum_{i=1}^{n} \sum_{j=1}^{d} [y_{ij} \ln a_{ij} + (1 - y_{ij}) \ln(1 - a_{ij})]$$

The learning of the network parameters is performed by means of stochastic gradient descent (Bottou, 2010) with a mini-batch size of 8 samples, considering the adaptive learning rate proposed by Zeiler (2012). This training stage consists of providing the SAE with examples of images and their corresponding ground-truth, that is, binary maps over the pixels that belong to musical symbols (see Fig. 4).

Once the SAE has been properly trained, removing staff lines from the image of a musical score consists of querying the image patch by patch. Each patch
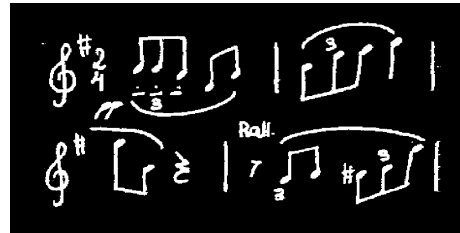
10

is passed through the SAE, which outputs the selection level assigned to each

₂₂₅ input pixel. Those pixels whose selection value exceeds a certain threshold are considered to belong to a musical symbol, whereas the others are discarded. Figure 5 illustrates this process.
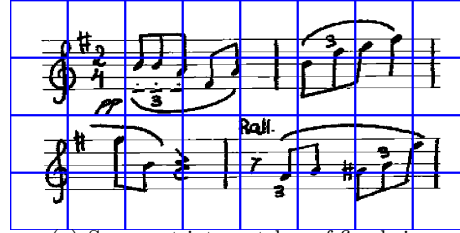
(a) Score cut into patches of fixed size
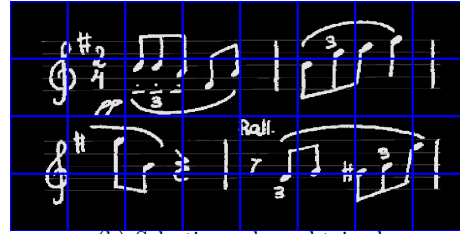
(b) Selection values obtained

(c) Score after thresholding

Figure 5: Example of staff-line removal task using an SAE. The input image is parsed patch by patch by means of the network. A selection value is predicted for each pixel in a patch (shown here as grayscale levels). Finally, a thresholding is applied in order to select the pixels that will eventually be maintained.

4. Experiments

This section presents the experiments carried out to evaluate the goodness of our proposal [1]. We took advantage of the *ICDAR / GREC 2013 Competition on Music Scores* (Fornés et al., 2013) staff-line removal contest by making use of the same dataset to allow reproducible research and future comparisons with our results.

This corpus contains trios consisting of a grayscale image of a score and its corresponding binarized version with and without staff lines (see Fig. 6). This dataset therefore provides readily-available data with which to train the model, along with testing data for evaluation. The corpora is organized into train and test sets, containing 4 000 and 2 000 samples, respectively.

(a) Grayscale score

(b) Binary score   (c) Ground-truth

Figure 6: Example of sample from the *GREC/ICDAR 2013 Staff-Line Removal Competition* dataset.

In our case, the train set is further subdivided into training and validation partitions. These partitions are distributed in 80 % and 20 % out of the total set of available training scores, respectively. Training data is used to perform

---

[1]For the sake of reproducible research, the code of the experiments is available at `http://github.com/ajgallego/staff-lines-removal` available under the conditions of the GNU General Public License version 3.

In order to evaluate the results, the *F-measure* (F-m) metric will be considered, following the guidelines of this contest to select the best method. Let TP be the number of true positives (symbol pixels correctly classified), FP be the number of false positives (symbol pixels incorrectly classified) and FN be the number of false negatives (staff-line pixels incorrectly classified). Therefore,

$$\text{F-m} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FN} + \text{FP}}$$

We first present the results obtained with the different topologies proposed for the SAE, with a study of the influence of certain hyper-parameters. We then compare our proposal with other methods for solving the same task that participated in the latest contest.

We also include further results concerning related issues, such as the robustness of the model with regard to the threshold. Before carrying out that analysis, we should state that the results shown below are obtained by assuming the best threshold for each case: 0.3 for binary images and 0.1 for the grayscale. The amount of data required to train the model successfully and the representational power of the model are also analyzed.

4.1. Hyper-parameter selection

This section presents the results obtained with different topologies of the SAE, namely the number of layers and input sizes. In order to reduce the search space, we have restricted ourselves to: i) considering symmetric models, that is, those with the same number of coding and decoding layers (from 1 to 3 for each stage); ii) considering only square images of sides 64, 128, 256, 384 and 512.

13

There are also a number of parameters to be tuned, such as the kernel size
₂₇₀ of the convolutions, the number of filters per layer and the batch size selected
during training. We have performed comprehensive experimentation in order to
tune these parameters by means of a grid search of:

- The number of filters per convolutional layer: 16, 32, 64, 96

- The kernel size of each convolutional layer: 3, 5, 7, 9, 11

₂₇₅ Since the number of configurations becomes huge, this first experiment fo-
cus on finding the optimal hyper-parameters only for the binary format of the
dataset. The idea of doing this hyper-parameter tuning only with binary images
is twofold: on the one hand, it is the case that has been traditionally studied
and, on the other hand, we do not want the grayscale results, the more complex
₂₈₀ scenario, depending on an excessive tuning of the model that best fits. Our
premise is that the goodness of our work lies in the proposed approach, and not
in selecting the optimum topology for every case. In addition, claiming that a
topology is the optimal one entails a very exhaustive search, and so we believe
that exploiting that avenue is not of actual interest in this work. We there-
₂₈₅ fore assume that the best configuration for this binary case will also achieve
competitive results in any other domain.

Table 1 shows the best result attained by each different SAE configuration
on the validation set. For the sake of readability, it merely reports the best
configuration as regards the number of filters and the kernel size of the con-
₂₉₀ volutional layers obtained for each combination of number of layers and input
size.

As an initial remark, we should state that our approach is relatively robust
to the different number of configurations, since all of the results consist of very
accurate figures. With regard to the window size, the results do not yield
₂₉₅ any magic number but the best figures appear to indicate that either smaller
or bigger window sizes would not significantly improve the results. On the
contrary, the number of layers has a higher impact on the performance, leading
to a variation in the F-m attained of up to 2 units.

14

# layers	Window size					Average
	64	128	256	384	512	
1	96.99	97.08	97.21	**98.40**	96.94	97.32
2	97.73	**98.70**	97.79	97.73	97.79	97.95
3	97.80	97.84	**99.13**	97.81	97.61	**98.01**

Table 1: F-m (%) attained by the different number of layers considered (rows) in combination with different values of input window size (columns) for binary images. The values in bold type highlight the best results in each row.

The model with 3 layers and 256×256 input patches performed best in ₃₀₀ this experimentation. Specifically, the complete configuration that reported the best accuracy comprises 96 filters per layer, and kernel sizes of 5. A technical description is given in Table 2. In the following, we shall assume that this configuration is a representative of our proposal for both binary and grayscale images.

₃₀₅ *4.2. Comparison with state-of-the-art*

Taking advantage of the aforementioned contest, we tested our method against state-of-the-art staff-line removal strategies. Since the conditions and participants are very different, the binary and grayscale format of the contest are analyzed separately.

₃₁₀ The test set provided is further divided into three subsets (TS1, TS2, and TS3) in order to measure the robustness of the participants with regard to the deformations applied to the scores: 3D distortions in TS1 (three types of distortions with 166, 167, and 167 samples; 500 scores in total), local noise in TS2 (two types of distortions with 250 samples each; 500 scores in total), ₃₁₅ and both 3D distortion and local noise in TS3 (6 specific distortions, equally distributed, as a result of the combination of the previous ones; 1 000 scores in total).

Table 3 shows the results obtained by the participants in the contest for the binary case. The main idea of each method was described in Sect. 2 — unlike

Input	Encoding	Decoding	Output
	Conv(96,5,5,ReLU)	Conv(96,5,5,ReLU)	
	MaxPool(2,2)	UpSamp(2,2)	
	Conv(96,5,5,ReLU)	Conv(96,5,5,ReLU)	
	MaxPool(2,2)	UpSamp(2,2)	
$[0, 255]^{256 \times 256}$	Conv(96,5,5,ReLU)	Conv(96,5,5,ReLU)	$[0, 1]^{256 \times 256}$
	MaxPool(2,2)	UpSamp(2,2)	
	Conv(96,5,5,ReLU)	Conv(96,5,5,ReLU)	
	MaxPool(2,2)	UpSamp(2,2)	
		Conv(1,5,5,Sigmoid)	

Table 2: Detailed description of the selected SAE architecture. Conv(f,h,w,a) stands for a convolution operator of f filters, with $h \times w$ pixel kernels with an a activation function; Max-Pool(h,w) stands for the max-pooling operator with a $w \times h$ kernel and stride; UpSamp(h,w) denotes an up-sampling operator of h rows and w columns; *ReLU* and *Sigmoid* denote Rectifier Linear Unit and Sigmoid activations, respectively.

³²⁰ *TAU*, which was a method specifically designed to participate in the contest. Those readers who require more detailed information about the participants are referred to the competition report (Visaniy et al., 2013), as some of the aforementioned strategies were slightly tuned for the contest. Moreover, we include the work of Calvo-Zaragoza et al. (2016) (Pixel), since their results were ³²⁵ obtained under the same conditions of the contest. The figures of our approach are those obtained by the selected configuration topology, based on the results depicted in the previous section.

As can be seen, most participants are able to achieve good performance figures, being some of them really close to the optimum. However, our approach ³³⁰ is able to improve on all the results obtained previously in all cases considered. This improvement is especially remarkable in TS3, when distortions are more aggressive. The comparison with Pixel method is also illustrative of the good-

	TS1	TS2	TS3	Whole
TAU (Visaniy et al., 2013)	85.72	81.72	82.29	83.01
NUS (Su et al., 2012)	69.85	96.25	67.43	75.24
NUASi-lin (Dalitz et al., 2008)	94.99	94.86	94.00	94.29
NUASi-skel (Dalitz et al., 2008)	94.25	93.80	92.92	93.34
LRDE (Géraud, 2014)	<u>97.73</u>	96.86	<u>96.98</u>	<u>97.14</u>
INESC (Dos Santos Cardoso et al., 2009)	89.29	97.72	88.52	91.01
Pixel (Calvo-Zaragoza et al., 2016)	94.10	<u>98.11</u>	94.00	95.04
Our approach	**99.11**	**99.36**	**99.03**	**99.13**

Table 3: F-m (%) comparison of the participants in the *ICDAR / GREC 2013* staff removal contest (binary format) and our approach based on SAE. The values in bold type represent the best result in each set, whereas the underlined values represent the best result without considering our model. The column *Whole* refers to the weighted average obtained for the three test sets.

ness of our proposal, since it demonstrates that the performance is not only achieved by using a supervised learning scheme (Pixel also does so) but because

₃₃₅ of the adequacy of the proposed model. Taking the test corpora as a whole, our approach is able to improve up to 2 points with respect to the best result achieved so far.

As mentioned previously, the dataset provided in the contest also contains a grayscale version of the scores. Our approach can easily be extended to deal

₃₄₀ with grayscale images with no further effort but to change the training data. In this case, only two of the methods submitted to the contest dealt with grayscale images: LRDE and INESC. Table 4 shows the results obtained by these participants, when compared to those obtained by our SAE, including again those of Pixel method.

₃₄₅ It is clear that the participants performance decreases remarkably, particularly as regards the INESC method. LRDE maintains a fair accuracy in TS1, but its performance is much worse in TS2 and TS3. Pixel method has a more robust behavior, and achieves similar figures regardless of the distortions applied to the images. With more room for improvement, our method achieves a

	TS1	TS2	TS3	Whole
LRDE (Géraud, 2014)	92.17	79.47	79.88	82.85
INESC (Dos Santos Cardoso et al., 2009)	38.50	52.11	38.87	42.09
Pixel (Calvo-Zaragoza et al., 2016)	<u>92.56</u>	<u>88.84</u>	<u>89.76</u>	<u>90.24</u>
Our approach	**99.14**	**99.34**	**98.94**	**99.09**

Table 4: F-m (%) comparison of the participants in the *ICDAR / GREC 2013* staff removal contest (grayscale format) and our approach based on SAE. The values in bold type represent the best result in each set, whereas the underlined values represent the best result without considering our model. The column *Whole* refers to the weighted average obtained for the three test sets.

performance far superior to the participants. It also undergoes a certain drop in accuracy with grayscale images but results still consist of very accurate figures, clearly outperforming the other methods. In this case, our approach is able to improve up to almost 10 points the state-of-the-art figure.

Results reported above only reflect the average performance of the methods. In order to minimize the possibility that the differences are due to chance variation, we perform a pairwise, non-parametric Wilcoxon signed-rank test Demsar (2006). We considered the 11 independent results (one per specific distortion applied) to perform these tests. It resulted in p-values below 0.01 in all pairwise comparisons between our approach and the other methods, in both binary and grayscale formats. Therefore, our approach proves to outperform the rest of the configurations with an alpha confidence level of 99%.

4.3. Analysis

The objective of this section is to analyze some of the characteristics of the SAE for the task of staff-line removal.

The quantitative results obtained by our method are very close to the optimum, especially for binarized images. Thus, it is interesting to look into the actual impact of such room for improvement. Figure 7 shows a qualitative example of the performance obtained by the SAE for a piece of image in both binary and grayscale formats. These examples depict an F-m of 99.06 % and 98.17

₃₇₀ %, respectively, therefore being good representatives of the results reported in the previous section. As can be observed, the differences between the outputs and the ground-truth are hardly perceptible. The mistakes might be related to pixels very close to the boundaries of symbols, which in any case does not seem to make a noticeable difference on the results.
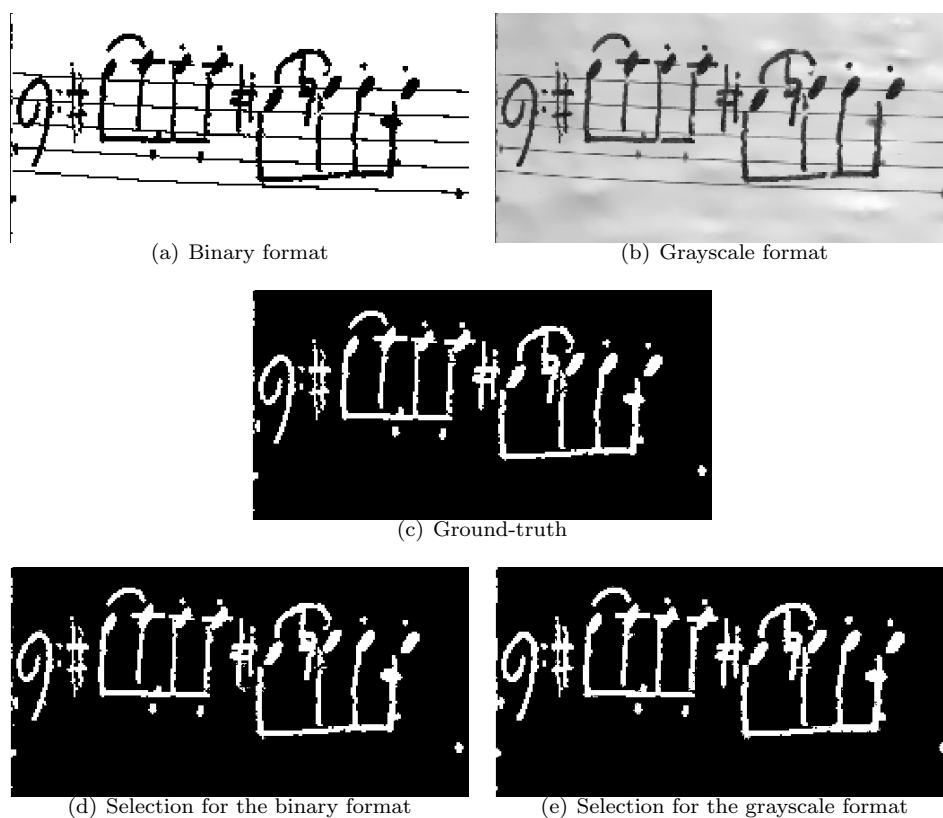

(a) Binary format

(b) Grayscale format

(c) Ground-truth

(d) Selection for the binary format

(e) Selection for the grayscale format

Figure 7: Example of a staff-line removal using the proposed SAE configuration for both binary and grayscale formats, depicting an F-m of 99.06 % and 98.17 %, respectively.

₃₇₅ At this point, it would be interesting to measure the impact of the accuracy of this staff-line removal in a functional OMR system, with respect to the accuracy obtained by considering previous works. Note that in musical notation it is important to be as precise as possible in this process: regions of FP (staff-line segments maintained) may involve the incorrect detection of small musical

19

symbols such as the *dot*, or changing the meaning of some notes (*quarter note* confused with an *eighth note*); similarly, regions of FN (symbol information removed) may make the system mislabel a *quarter note* as a *half note*. Taking into account that music notation must fulfill strong grammatical constraints, these hypothetical isolated mistakes may cause many errors in other regions of the score. Unfortunately, there are no reproducible strategies for complete OMR workflow, and so carrying out this experiment would imply a study beyond the scope of this paper. However, we place this issue at future work considerations.

Furthermore, having demonstrated the accuracy of the approach, it is important to stress its principal differences from conventional methods for staff-line removal. As stated above, one interesting property is that the SAE follows a supervised learning approach, so its extension as regards dealing with any type of input image is feasible. This applies not only in the case of a different nature of the image —such as RGB images— but also in that of different types of features of the score, different notational styles, heterogeneous document conditions and so on. It is true, however, that the approach cannot be immediately applied to any domain but our claim is that it is much easier to set up the SAE environment (basically, training data and some parameter tweaking) than developing a completely new staff-line removal strategy.

Pixel method is also based on supervised learning but there are two important differences. The first is that our method achieves significatively better results, while the second, which is indeed the most relevant from a practical point of view, is that Pixel method is a computationally expensive technique, since it has to classify every single pixel in the image. In this respect, once the SAE has been trained, the computation time needed to process a score is in the order of seconds, while that of Pixel method is in the order of hours.

In the following sections, we delve into interesting aspects of the proposed model. We first study the influence of the threshold used to convert SAE predictions into the binary selectional output. An incremental learning scenario is also considered to check the number of samples needed for the model to learn the task. Finally, we show the representational capabilities of the SAE by analyzing

20

the intermediate codification of the input patch.

4.3.1. Threshold

As stated in Section 3, the SAE predicts a value for each input feature, which can be understood as the level of selection. In this respect, for the results shown above it was assumed that a threshold equal to 0.3 (binary images) and 0.1 (grayscale images) would discriminate between whether or not the feature was eventually activated. Here we report the reason why these values were chosen and the real impact that other configurations have.

Figure 8 shows the F-m obtained for different threshold values — within the range of $[0, 1]$ — considering the binary and grayscale format of the images. The values obtained by the best state-of-the-art algorithms have also been included for the sake of comparison.

It can be observed that, although the thresholds considered are effectively those that obtain the best performances, any other threshold would have outperformed the state-of-the-art strategies with a fair margin. These figures reinforce the robustness of the activations predicted by the SAE. These values tend to be close to either 0 or 1, thereby decreasing the importance of the threshold.

4.3.2. Training set size

We shall now focus on assessing the impact of the amount of data used to train the model on the performance figures. To this end, an additional experiment was performed in which the number of training scores was iteratively increased.

The results of this study are shown in Fig. 9. In the case of binary images, in which the variability is much lower, the SAE is able to perform very competitively with a relatively small number of scores. The model outperforms the state-of-the-art result with 50 scores. On the contrary, the case of grayscale images is more complex, and more samples are needed to reach stable results. Note, however, that the state-of-the-art is reached with a few number of training scores.

21

(a) Binary format
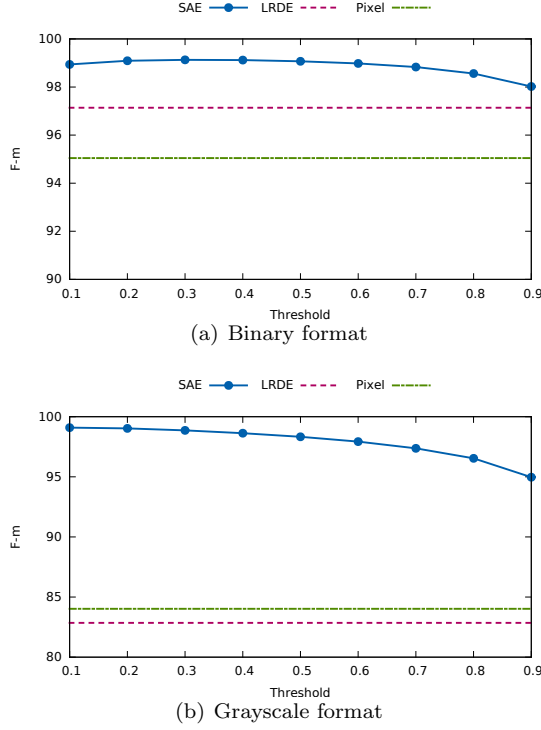

(b) Grayscale format

Figure 8: Influence of the threshold parameter on the performance of the SAE. Performance of state-of-the-art strategies are also included for a better comparison.

⁴⁴⁰ Furthermore, it should be emphasized that, in both cases, the curve shows an upward tendency when reaching the total number of training scores. It seems that a greater number of training samples could still improve the results obtained, especially in the case of grayscale.

4.3.3. Intermediate representation

⁴⁴⁵ As a final analysis, this section studies the type of intermediate representation that the SAE encodes. In the case of the model that we have determined is the most appropriate for this task (see Section 4.1), the intermediate representation comprises $96 \times 32 \times 32$ features. In other words, it codifies 96 images of 32×32.

⁴⁵⁰ A full example of intermediate representation can be seen in Fig. 10. In

22

(a) Binary format

(b) Grayscale format
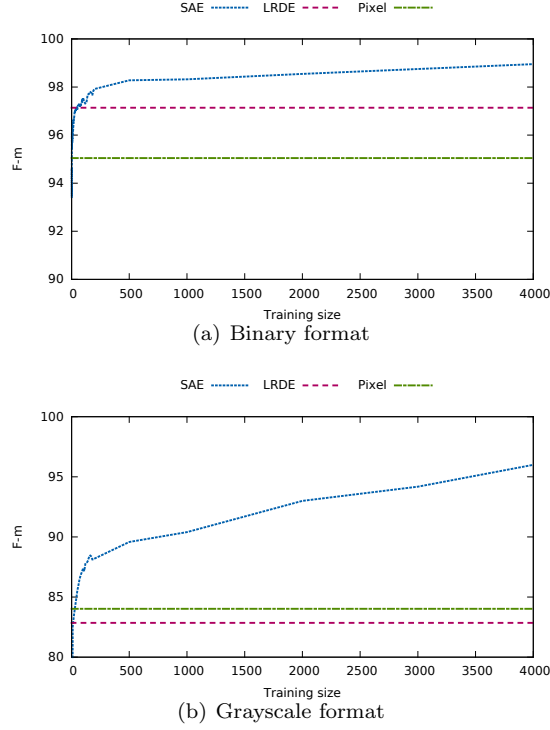
Figure 9: Performance of the SAE with regard to the amount of training scores used to train. Performance of state-of-the-art strategies are also included for a better comparison.

our grid search, which was carried out to determine the best model parameters, the number of intermediate codes did not lead to an excessive variation in the performance figures, which is why many of the intermediate representations are quite similar: there is a high redundancy of information in the case of 96 codes, signifying that similar results could be obtained with a fewer number of them. The actual utility of all the codes might only arise in the most complex cases.

Furthermore, in order to compare the difference between binary and grayscale input images, Table 5 shows examples of the intermediate codifications obtained from both formats (enlarged for visualization). It is interesting to note that many of these intermediate representations are also quite similar, regardless of the type of input image. This could evidence that the SAE is actually learning a good internal representation to perform this task, discarding the information

23
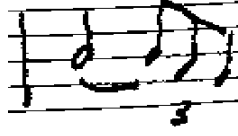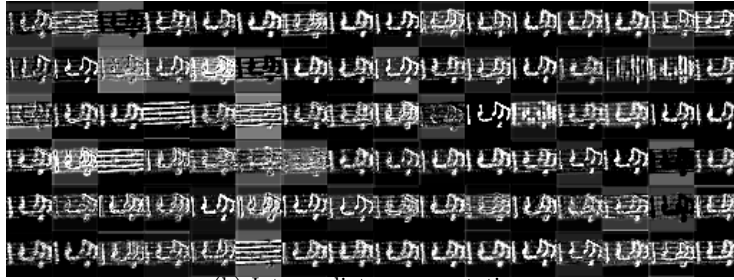
(a) Input patch

(b) Intermediate representations

Figure 10: Illustration of the 96 intermediate representations for the depicted input and output patches.

that refers to the specific characteristics of the input image.

4.4. Experiment with old documents

⁴⁶⁵ A new experiment is described in this section with the aim at verifying the adaptability of the approach model to a different type of document images. It is obvious that, in this case, data-driven strategies have advantages because they are provided with information of the specific domain on which they are going to be applied. Traditional methods for staff-line removal have not taken into ⁴⁷⁰ account the great heterogeneity that can be found in musical documents, thus leading to solutions that are not generalizable. We want to show here that the SAE entails a more adaptable approach, and it also behaves better than other supervised learning approaches.

For this experiment we use a set of 20 staff sections from sacred music com-⁴⁷⁵ posed during the second half of the XVIIth century. These compositions were handwritten in a music book by a copyist of that time. In addition, they do not depict common modern notation, but the so-called mensural notation. Since

24
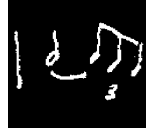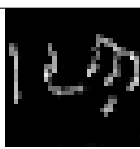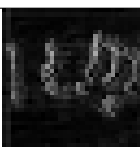
Input	Intermediate representations (enlarged)	Output

Table 5: Example of intermediate representation of the SAE.

music was intended to be sung, the sections depict both music notation and accompanying text (lyrics), which might hinder the staff-line removal algorithms.

We have manually created a binary version of the considered staff sections, which are originally depicted in grayscale, as well as their corresponding ground-truth (without staff lines). An example of this corpus is illustrated in Table 6.
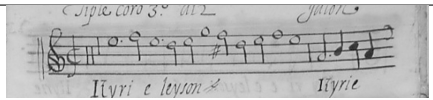
Source
Binary
Ground truth

Table 6: Examples of the corpus created for staff-line removal on mensural notation manuscripts.

Unlike the previous case, for which there exist a large dataset with thousands of images to train the models, this experiment shows a more real scenario. In

₄₈₅ such a situation, it cannot be assumed that a large corpus can be effortlessly obtained, but the network must learn with a limited amount of examples. However, we resort to fine-tuning to alleviate this situation. That is, a pre-trained model for other types of scores (*ie.* that obtained in the previous experiment) is initially considered, and then examples of the new domain are used to re-
₄₉₀ estimate the parameters of the model. In addition, the extracted patches from the training images are not totally disjoint but overlapping is considered in order to have a larger number of training examples from the same number of labeled documents.

In this experiment, we first focus on selecting the most appropriate strategy
₄₉₅ to approach a new domain (mensural notation). The first option is to directly use the pre-trained network obtained from the previous experiment with modern notation. However, we also measure the performance obtained assuming that we have a limited number of tagged data of the new domain. Thus, we include in the comparison the performance of a trained network from scratch with these
₅₀₀ new data, as well as starting from the pre-trained network and performing a fine tuning process.

To take the most out of the available data, we have established a 5-fold cross-validation scheme. That is, for each fold, 16 images are considered for training (20% of which are used to monitor the training and prevent over-fitting) and
₅₀₅ 4 images are used to measure the performance. The average results of this experiment can be checked in Table 7.

In this case we can observe relevant differences between binary and grayscale formats. In the formed, the network trained with other types of scores is quite reliable—even better than training it from scratch with the new data—as the
₅₁₀ staff lines are similar to those depicted in the previous corpora. However, it is appreciated that adding data from the new domain allows boosting the recognition. In the grayscale format, it is observed that training with domain data, even with a limited number of examples, is more beneficial than directly consider the pre-trained network with data of another type of score. The difference with
₅₁₅ the binary case is that here the model has to deal also with the background of

26

Method	Binary	Grayscale
Pre-trained	96.16	84.05
Trained from scratch	95.05	91.45
Pre-trained + fine-tuning	**97.98**	**95.71**

Table 7: F-m (%) comparison among the different strategies for training the SAE in a new domain (mensural notation). Results report average performance considering a 5-fold cross validation experiment in both binary and grayscale format. Values in bold represent the best average accuracy in each set.

the score, which presents relevant differences with respect to the background of the previous corpus. Nevertheless, it is also reported that the best performance is achieved when combining a pre-trained network with the new data.

Finally, we compare the best case obtained by our model (that is, starting ₅₂₀ with a pre-trained model and fine-tuning) with the algorithms previously proposed. In particular, we consider the 3 methods that obtained the best result in the benchmark established by the aforementioned staff-line removal contest: LRDE, INESC and Pixel. The first two were specially designed for binary images, and so we only consider their most favorable case. Furthermore, Pixel is ₅₂₅ easily usable in both contexts. Thus, the results of this comparative study can be consulted in Table 8.

It can be observed that the SAE is able to outperform the results obtained with other approaches, even with a greater difference than in the previous experiment, in both binary and grayscale. Although the supervised approach is ₅₃₀ clearly an advantage in this context, it is also reported that our approach noticeably improves the results obtained by Pixel. An illustrative detail of these results is given in Fig. 11.

A statistical significance test is performed again, considering each complete document as an independent sample. These tests resulted in p-values below ₅₃₅ 0.01 when comparing our method against the rest of the strategies, implying therefore that our method significantly improves their performance with an

27

Method	Binary	Grayscale
LRDE	92.81	—
INESC	90.89	—
Pixel	91.24	86.64
SAE	**97.98**	**95.71**

Table 8: F-m (%) comparison among LRDE, INESC, Pixel, and SAE methods for the staff-line removal over mensural notation manuscripts. Results report average performance considering a 5-fold cross validation experiment in both binary and grayscale format. Values in bold represent the best average accuracy in each set. A dash mark (—) is used when the method in the row is not applicable.

alpha confidence level of 99%.

5. Conclusions

This work has studied the removal of staff lines from musical scores by considering a machine learning approach. Our proposal consists of using a new type of model called the Selectional Auto-Encoder (SAE), a convolutional neural network that learns which characteristics of the input should be selected by means of coding and decoding stages. In the context of the problem to be addressed, the model is trained to select only those pixels of the input that belong to a musical symbol. The process of removing the staff lines can therefore be solved, once the model is trained appropriately, by iterating the image of a score patch by patch. An activation in the range of $[0, 1]$ is obtained for each pixel, indicating the level of selection predicted. Those pixels whose activation surpass a certain threshold are considered to be part of a musical symbol, and are otherwise discarded.

Our comprehensive experimentation on a standard dataset has demonstrated the goodness and robustness of the approach. Regardless of the specific features of the chosen model — such as the size of the input window or the depth of the network — the results obtained were competitive. The model that obtained the
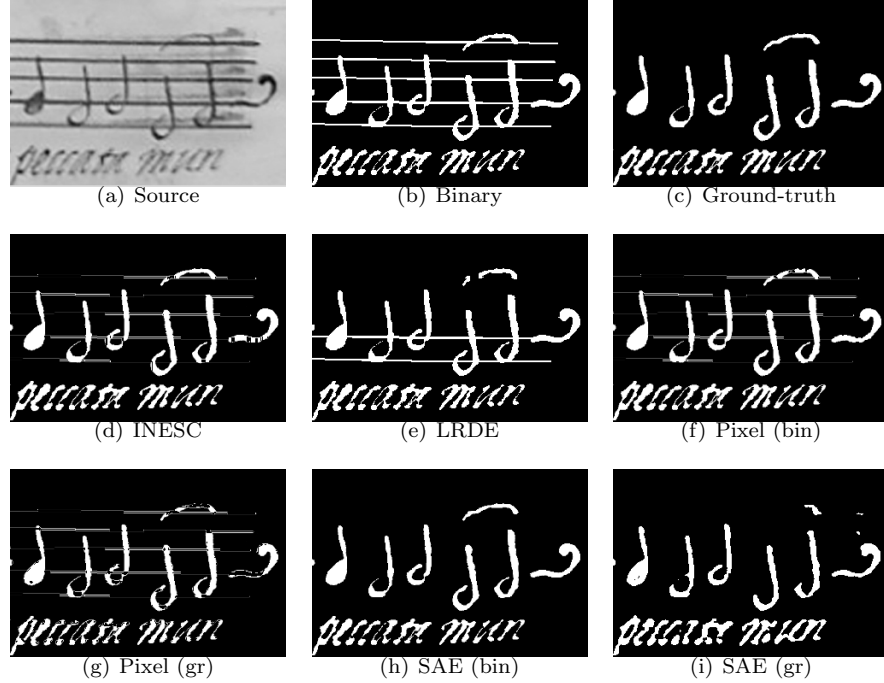
28

Figure 11: Qualitative detail of the performance achieved by the different staff-line removal strategies for mensural notation from old manuscripts. Grayscale format (source), manual binarization, and ground-truth are also included for a better illustration.

best figures was specifically that which takes a window of size 256×256, with 3 convolutional layers per stage, 96 filters per layer and a 5×5 convolution kernel per filter. When compared to other algorithms proposed for the same task, the SAE performs significatively better for both binary and gray input images. The goodness of our approach is more evident in the latter case (up to more than 10 points of improvement in F-m).

We have also studied the behavior of our model with regard to the chosen threshold, showing that this value does not have a particular impact on the results. However, it has been determined that a value of 0.3 is the most appropriate for dealing with binary inputs, and 0.1 for grayscale. An incremental study has also been carried out, and we have observed that the model attains a competitive performance with few samples. However, in the case of grayscale

29

images, the best values are only achieved with a large number of examples. Finally, we have shown the intermediate representations that the SAE learns, regarding which two main observations can be made: these representations have

⁵⁷⁰ a high redundancy of information, which could mean that the whole capacity of the model is only needed in the most complex cases; the intermediate representations for the binary and grayscale cases are quite similar, which could indicate that the SAE is actually learning how to attain a *deep* representation of the task, independently of the characteristics of the input image.

⁵⁷⁵   As prospects for future work, the intention is twofold. On the one hand, we intend to include our staff-line removal process in a functional OMR system so as to study its impact in a goal-directed way. We are especially interested in measuring the final performance of the system as regards the staff-line removal accuracy. On the other hand, we are interested in learning the task with a

⁵⁸⁰ limited number of labeled samples. Note that the reference corpus for the problem of staff-line removal is expensive to obtain. The idea is, therefore, to follow semi-supervised approaches in which the task can be learned in an unsupervised manner and a fine-tuning process with a small labeled corpus is carried out. It would be of great interest to carry out studies by increasing the variability of

⁵⁸⁵ the input images. Musical scores can be highly heterogeneous, and we thus wish to obtain a model that can adapt to any type of them. Our idea is, therefore, to develop models that can process scores of a very different style from those seen during training by means of Transfer Learning (Pan & Yang, 2010) or Domain Adaptation (Patel et al., 2015) strategies.

References

Bainbridge, D., & Bell, T. C. (1997). Dealing with superimposed objects in optical music recognition. In *Proceedings of the 6th International Conference on Image Processing and Its Applications* (pp. 756–760).

Bengio, Y., Yao, L., Alain, G., & Vincent, P. (2013). Generalized denoising auto-encoders as generative models. In *Advances in Neural Information Processing Systems* (pp. 899–907).

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010* (pp. 177–186). Springer.

Calvo-Zaragoza, J., Barbancho, I., Tardón, L. J., & Barbancho, A. M. (2015). Avoiding staff removal stage in optical music recognition: application to scores written in white mensural notation. *Pattern Anal. Appl.*, *18*, 933–943.

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

Carter, N., & Bacon, R. (1992). Automatic recognition of printed music. In H. Baird, H. Bunke, & K. Yamamot (Eds.), *Structured Document Image Analysis* (pp. 454–465). Springer.

Choudhury, G. S., Droetboom, M., DiLauro, T., Fujinaga, I., & Harrington, B. (2000). Optical music recognition system within a large-scale digitization project. In *ISMIR 2000, 1st International Symposium on Music Information Retrieval, Plymouth, Massachusetts, USA, October 23-25, 2000, Proceedings*.

Dalitz, C., Droettboom, M., Pranzas, B., & Fujinaga, I. (2008). A comparative study of staff removal algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, *30*, 753–766.

Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, 1–30.

Deng, L., Seltzer, M. L., Yu, D., Acero, A., Mohamed, A., & Hinton, G. E. (2010). Binary coding of speech spectrograms using a deep auto-encoder. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010* (pp. 1692–1695).

Dos Santos Cardoso, J., Capela, A., Rebelo, A., Guedes, C., & Pinto da Costa, J. (2009). Staff Detection with Stable Paths. *IEEE Trans. Pattern Anal. Mach. Intell.*, *31*, 1134–1139.

Dutta, A., Pal, U., Fornes, A., & Llados, J. (2010). An Efficient Staff Removal Approach from Printed Musical Documents. In *2010 20th International Conference on Pattern Recognition (ICPR)* (pp. 1965–1968).

Fornés, A., Kieu, V. C., Visani, M., Journet, N., & Dutta, A. (2013). The IC-DAR/GREC 2013 Music Scores Competition: Staff Removal. In *10th International Workshop on Graphics Recognition, Current Trends and Challenges GREC 2013, Bethlehem, PA, USA, August 20-21, 2013, Revised Selected Papers* (pp. 207–220).

Fujinaga, I. (2005). Staff detection and removal. In S. George (Ed.), *Visual Perception of Music Notation* (pp. 1–39). Hershey, PA: Idea Group Inc.

Fujinaga, I., Hankinson, A., & Cumming, J. E. (2014). Introduction to SIMSSA (single interface for music score searching and analysis). In *Proceedings of the 1st International Workshop on Digital Libraries for Musicology, DLfM@JCDL 2014, London, United Kingdom, September 12, 2014* (pp. 1–3).

Géraud, T. (2014). A Morphological Method for Music Score Staff Removal. In *Proceedings of the 21st International Conference on Image Processing (ICIP)* (pp. 2599–2603). Paris, France.

Hankinson, A., Burgoyne, J. A., Vigliensoni, G., & Fujinaga, I. (2012). Creating a large-scale searchable digital collection from printed music materials. In

Proceedings of the 21st World Wide Web Conference, WWW 2012, Lyon, France, April 16-20, 2012 (Companion Volume) (pp. 903–908).

Hinton, G. E., & Zemel, R. S. (1994). Autoencoders, minimum description length and helmholtz free energy. In *Advances in Neural Information Processing Systems* (pp. 3–10).

Lauly, S., Larochelle, H., Khapra, M., Ravindran, B., Raykar, V. C., & Saha, A. (2014). An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems* (pp. 1853–1861).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.

Martin, P., & Bellissant, C. (1991). Low-level analysis of music drawing images. In *First International Conference on Document Analysis and Recognition* (pp. 417–425).

Ng, K. (2001). Music manuscript tracing. In *International Workshop on Graphics Recognition* (pp. 330–342). Springer.

Ntirogiannis, K., Gatos, B., & Pratikakis, I. (2014). ICFHR2014 competition on handwritten document image binarization (H-DIBCO 2014). In *14th International Conference on Frontiers in Handwriting Recognition, ICFHR 2014, Crete, Greece, September 1-4, 2014* (pp. 809–813).

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, *22*, 1345–1359.

Patel, V. M., Gopalan, R., Li, R., & Chellappa, R. (2015). Visual Domain Adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, *32*, 53–69.

Ramirez, C., & Ohya, J. (2014). Automatic recognition of square notation symbols in western plainchant manuscripts. *Journal of New Music Research*, *43*, 390–399.

Randriamahefa, R., Cocquerez, J. P., Fluhr, C., Pepin, F., & Philipp, S. (1993). Printed music recognition. In *Document Analysis and Recognition, 1993., Proceedings of the Second International Conference on* (pp. 898–901). IEEE.

Raphael, C., & Wang, J. (2011). New approaches to optical music recognition. In *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, Miami, Florida, USA, October 24-28, 2011* (pp. 305–310).

Rebelo, A., Capela, G., & Cardoso, J. S. (2010). Optical recognition of music symbols - A comparative study. *International Journal on Document Analysis and Recognition*, *13*, 19–31.

Rebelo, A., & Cardoso, J. (2013). Staff Line Detection and Removal in the Grayscale Domain. In *2013 12th International Conference on Document Analysis and Recognition (ICDAR)* (pp. 57–61).

Rebelo, A., Fujinaga, I., Paszkiewicz, F., Marçal, A. R. S., Guedes, C., & Cardoso, J. S. (2012). Optical music recognition: state-of-the-art and open issues. *International Journal of Multimedia Information Retrieval*, *1*, 173–190.

Roach, J., & Tatem, J. (1988). Using domain knowledge in low-level visual processing to interpret handwritten music: an experiment. *Pattern recognition*, *21*, 33–44.

Su, B., Lu, S., Pal, U., & Tan, C. (2012). An Effective Staff Detection and Removal Technique for Musical Documents. In *2012 10th IAPR International Workshop on Document Analysis Systems (DAS)* (pp. 160–164).

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, *11*, 3371–3408.

Visaniy, M., Kieu, V., Fornes, A., & Journet, N. (2013). ICDAR 2013 Music Scores Competition: Staff Removal. In *2013 12th International Conference on Document Analysis and Recognition (ICDAR)* (pp. 1407–1411).

Wang, W., Huang, Y., Wang, Y., & Wang, L. (2014). Generalized autoencoder: A neural network framework for dimensionality reduction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (pp. 490–497).

Wen, C., Rebelo, A., Zhang, J., & Cardoso, J. S. (2015). A new optical music recognition system based on combined neural network. *Pattern Recognition Letters*, *58*, 1–7.

Zeiler, M. D. (2012). ADADELTA: an adaptive learning rate method. *CoRR*, *abs/1212.5701*.