Machine Vision and Applications manuscript No.
(will be inserted by the editor)

Staff-line detection and removal using a Convolutional Neural Network

Jorge Calvo-Zaragoza · Antonio Pertusa · Jose Oncina

Received: date / Accepted: date

Abstract Staff-lines removal is an important preprocessing stage for most Optical Music Recognition systems. Common procedures to solve this task involve image processing techniques. In contrast to these traditional methods based on hand-engineered transformations, the problem can also be approached as a classification task in which each pixel is labeled as either *staff* or *symbol*, so that only those that belong to symbols are kept in the image. In order to perform this classification we propose the use of Convolutional Neural Networks, which have demonstrated an outstanding performance in image retrieval tasks. The initial features of each pixel consist of a square patch from the input image centered at that pixel. The proposed network is trained by using a dataset which contains pairs of scores with and without the staff lines. Our results in both binary and grayscale images show that the proposed technique is very accurate, outperforming both other classifiers and the state-of-the-art strategies considered. In addition, several advantages of the presented methodology with respect to traditional procedures proposed so far are discussed.

Keywords Music staff-lines removal · Optical Music Recognition · Pixel classification · Convolutional Neural Networks

1 Introduction

Music is an important part of the cultural heritage, which represents a key element for understanding the social and artistic trends of a specific period of history. That is why a large number of music documents have been carefully preserved over the centuries, scattered across cathedrals, museums and historical archives.

Massive digitization is an indispensable step for the preservation of these documents, and it sets the basis towards the development of tools that would facilitate the access, search and study of these sources. In addition, it would open several opportunities to apply Music Information Retrieval algorithms [26], which may be of great interest for the musicology community since these algorithms might be able to go beyond what a human can achieve after years of study.

Manual transcription of music, however, is an expensive task because it has to be carried out by music experts. Moreover, the complexity of music notation inevitably leads to burdensome software for music score edition, in which a long transcription process becomes tedious and prone to errors. As a consequence, the de-

This work was supported by the Spanish Ministerio de Educación, Cultura y Deporte through a FPU Fellowship (Ref. AP2012–0939), the Spanish Ministerio de Economía y Competitividad through Project TIMuL (No. TIN2013-48152-C2-1-R supported by EU FEDER funds) and the Instituto Universitario de Investigación Informática (IUII) from the University of Alicante. Authors would like to thank the anonymous reviewers for their constructive comments to improve the paper quality.

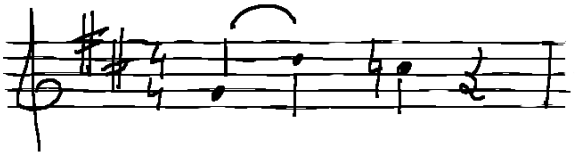
Jorge Calvo-Zaragoza
Department of Software and Computing Systems, University of Alicante
Carretera San Vicente del Raspeig s/n, 03690 Alicante, Spain
Tel.: +349-65-903772
E-mail: jcalvo@dlsi.ua.es

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

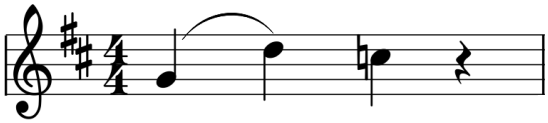
Jose Oncina
Department of Software and Computing Systems, University of Alicante
Carretera San Vicente del Raspeig s/n, 03690 Alicante, Spain

velopment of automatic transcription systems is gaining importance over the last years.

Optical Music Recognition (OMR) is the field of computer science devoted to understanding the musical information contained in the image of a music score [2]. The process aims at importing a scanned music score and exporting its musical content into some machine-readable format (Fig. 1).



(a) Example of input score for an OMR system



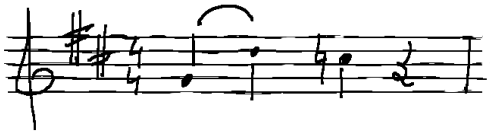
(b) Symbolic representation of the input score

Fig. 1 The task of Optical Music Recognition (OMR) is to analyze an image containing a music score to export its musical content into some machine-readable format.

From a morphological point of view, music hardly has what we might consider low-level entities — like characters in text or phonemes in speech — but rather isolated symbols. Consequently, the recognition of musical documents might be seen similar to the task of Optical Character Recognition.

However, the complexity of musical notation is higher than other similar domains (*eg.*, text), reflected in many ways such as the fact of finding symbols with a double nature (rhythm and harmony) and the possibility of finding several symbols sharing the same vertical position (chords, dynamics, ligatures, etc.). These issues lead to discard continuous recognition models and focus on segmentation plus classification approaches. In this sense, although research has been conducted on the recognition of isolated music symbols [20], OMR comprises a greater challenge since their detection and segmentation is not a trivial matter. One of the important aspects to consider is the presence of the *staff*, the set of five parallel lines that appear in sheet music to indicate the pitch of the notes. These lines are necessary for human readability, yet they complicate the automatic isolation of music symbols and removing them is a key step in the resolution of the OMR task.

Only a few works have taken advantage of specific features of printed notation to approach the problem maintaining the staff lines [18,3]. However, the established OMR pipeline includes the staff-line detection and removal task [21] to remove the staff lines while keeping as much as possible the symbol information (Fig. 2).



(a) Example of input score for an OMR system

(b) Input score after staff lines removal

Fig. 2 Example of a perfect staff lines removal process.

In this paper we propose the use of convolutional neural networks to solve this task. This approach is inspired by the work of Calvo Zaragoza et al. [4], in which this process is carried out as a classification task at pixel level. We show this time that the use of this kind of networks is able to outperform specific strategies based on image processing, as well as conventional classifiers. In addition, for the first time we extend the applicability of this strategy to deal with grayscale images.

The rest of the paper is organized as follows: Section 2 puts into context our contribution; Section 3 describes our approach; a comprehensive experimentation is showed in Section 4; Section 5 analyzes the pros and cons of this strategy, according to results obtained; finally, Section 6 concludes the present work and highlights some interesting lines of future research.

2 Background

Although staff lines detection and removal may be seen as a simple task, it is often difficult to get accurate results. This is mainly due to sheet deformations such as discontinuities, skewing or paper degradation (especially in ancient documents). In addition, musical documents are very heterogeneous so it has been extremely difficult to develop methods that are able to work on any kind of scores. A comprehensive review of the first

attempts considered for this task can be consulted in the work of Dalitz et al. [6]. More recently, however, many other methods has been proposed.

Dos Santos Cardoso et al. [7] proposed a method that considers the staff-lines as connecting paths between the two margins of the score. Then, the score is modeled as a graph so that staff detection is solved as a maximization problem. This strategy was improved and extended to be used on gray-scale scores [19]; Dutta et al. [8] developed a method that considers the staff line segment as a horizontal connection of vertical black runs with uniform height, which were validated using neighboring properties; in the work of Piatkowska et al. [17], a Swarm Intelligence algorithm is applied to detect the staff line patterns; Su et al. [24] started estimating properties of the staves like height and space. Then, they tried to predict the direction of the lines and fit an approximate staff, which is posteriorly adjusted; Geraud [11] developed a method that entails a series of morphological operators directly applied to the image of the score to remove staff lines; Montagner et al. [16] proposed to learn image operators, following the work of Hirata [13], whose combination is able to remove staff lines. On the other hand, some studies addressed the whole OMR problem by developing their own, case-directed staff removal process [22,25].

As presented above, several procedures for staff detection and removal have been proposed in the literature. Although in many cases most of them show a very good performance, they are far from optimal when the style of the score changes, as they rely on characteristics that are particular to a given style. Conversely, a data-driven strategy for staff removal has been recently considered, which consists of a classifier that discriminates if an *ink* pixel belongs to a symbol or to a staff line [4]. A supervised learning algorithm can be trained using the neighboring pixels as feature vector. Then, the foreground pixels of the image are queried so that those classified as *staff* are removed. Although this strategy achieved a fair performance, it reported room for improvement regarding the accuracy since it did not reached a state-of-the-art performance. In the presented study we propose to extend and further evaluate this approach by using more appropriate classification techniques.

In recent years, Convolutional Neural Networks (CNN) have shown a great ability in classification tasks when dealing with images, and generally with signals [5]. This study aims at using this kind of networks for pixel classification in the context of staff lines removal. One of the main advantages of these networks is that they are able to learn the intrinsic representation of the input data and, therefore, there is no need of hand-crafted feature extraction.

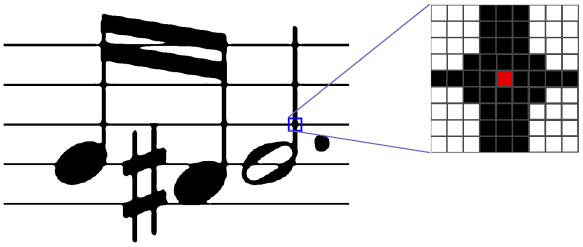


Fig. 3 Example of feature extraction with a square patch. The center pixel (marked in red) is the one to be classified. The pixel of this example must classified as *symbol* because if belongs to the stem of a half note.

3 Staff-Line Removal with Convolutional Neural Networks

As mentioned above, the staff-lines removal problem is tackled here from the point of view of a classification task. Basically, the strategy is to query each pixel of the image to either keep it because it belongs to a musical symbol or remove it because it belongs to a staff line. To do this, we use representative data of each pixel of interest and a CNN trained to distinguish between these two classes.

3.1 Input data

Although OMR systems have to deal with color pages, a typical image preprocessing step is to binarize them [21]. In this study we do not need to assume such condition. However, we consider two different scenarios: one in which the images have been previously binarized and another in which the images are presented in grayscale format. As we will see later, our approach is equally applicable regardless of the scenario chosen.

The neural networks to be trained must distinguish if a foreground pixel belongs to a staff or to a symbol. 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. 3).

Figure 4 shows several examples of input data for each class (symbol and staff). For the sake of visualization, these examples show 28×28 windows. Note that the label of each patch depends on the pixel located in the center.

It is clear that different sizes of the surrounding region could be taken into account so this parameter will be experimentally evaluated.

(a) Samples of *staff* class

(b) Samples of *symbol* class

Fig. 4 Examples of samples from both *staff* and *symbol* classes if 28×28 windows are considered.

In order to obtain a proper training set, this feature extraction process is applied to the foreground pixels of a dataset of scores labeled with and without staff lines. Given a foreground pixel in the position (i, j), a square patch is extracted from the score that contains staff lines (*i.e.,* the original one), whereas the value in the position (i, j) of the score without staff lines is used to obtain the actual label that can be either *staff* or *symbol*. No further feature extraction is performed on this portion of image because this task is expected to be assumed by the neural network.

3.2 Convolutional Neural Networks

The main advantage of using machine learning for the staff-line detection and removal problem lies in its ability to generalize, in comparison to systems based on hand-crafted image processing strategies. While the latter focus on singular aspects of the documents to be analyzed—being therefore very difficult to adapt them to other types of documents of different epoch, notation, or style—techniques based on supervised learning only need labeled examples of new documents to generate a model adapted to the new environment.

Convolutional Neural Networks (CNN) significantly outperform traditional techniques in a wide range of image recognition tasks [15]. These networks take advantage of local connections, shared weights, pooling, and many connected layers that eventually learn a data representation suitable for the task at hand.

Since the topology of a CNN can be quite varied, we decided to carry out an exhaustive search of a suitable configuration for the problem at issue. However, in order to reduce the search space, we have designed a CNN in which i) each convolutional layer consists of 3×3 filters and 2×2 max-pooling (VGG-alike [23]); ii) only one fully-connected layer is added at the end of the network, with 64 units and 50 % of dropout; and iii) only square input images are considered. The rest of the configuration parameters (size of the input layer, depth of the network, number of filters per layer and type of activations) will be assessed empirically.

The learning of the network weights is performed by means of stochastic gradient descent [1] with a batch size of 32, considering the adaptive learning rate proposed in [28] (default parameterization) and the *cross-entropy* loss function.

Once the CNN has learned how to distinguish between *staff* and *symbol* patches, it can be used to remove the staff lines of any input image. To do so, each pixel of that image can be forwarded with its patch through the network, and those pixels classified as *staff* will eventually be removed.

4 Experiments

Taking advantage of the *ICDAR 2013 Competition on Music Scores* [10] staff removal contest, we follow its same experimental set-up to assure a fair comparison with other studies and provide reproducible research. This corpus contains pairs of scores with and without staff lines. Therefore, this dataset provides available data for training the network, as well as testing data for evaluating our approach. In addition, this contest will allows us to test our method against state-of-the-art staff removal strategies.

The corpora used in this contest is organized in train and test sets, with $4,000$ and $2,000$ samples respectively. This dataset assumes that the scale has been normalized so that the space between lines is of 26 pixels. The test set is further divided into three subsets (TS1, TS2, and TS3) based on the deformations applied to the scores: 3D distortions in TS1 (500 scores), local noise in TS2 (500 scores), and both 3D distortion and local noise in TS3 (1000 scores). Each sample consists of an image of a handwritten score with its corresponding ground-truth (the score without staves). On average, the number of foregrounds pixels per score is around $500,000$, with $200,000$ staff pixels.

The train set is used to train the CNN that classifies between staff and symbol pixels. A part of this set is used as validation data to select the most appropriate epoch to stop the learning process and prevent overfitting.

The test corpora, which is not seen during training, is used to measure the performance. As in the competition, the performance metric is the *F-measure* or F_1

score:

$$F_1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

where TP, FP and FN stand for true positives (symbol pixels classified as symbol), false positives (staff pixels classified as symbol) and false negatives (symbol pixels classified as staff), respectively.

4.1 Network tuning

As mentioned before, some of the configuration parameters of the network are going to be adjusted experimentally. In order to clarify which configuration constitutes the most suitable topology, we carry out an experiment using only the training set of the contest. Given that there are enough data, 400 000 samples are used to perform the training of the network, chosen randomly among all the pages of the training set, whereas the rest are used as validation set.

There are a number of parameters to be tuned in the aforementioned CNN model. For this, we have performed a comprehensive experimentation in order to tune these parameters by means of a *grid search* over:

- Size of input patches: odd values (because of the central pixel) from $7x7$ to $29x29$.
- Depth of the network (number of convolutional plus max-pooling blocks): 1, 2, 3.
- Number of filters per convolutional layer: 16, 32, 64.
- Activation function: Rectified Linear Unit (ReLU) and hyperbolic tangent (tanh).

Note that some combinations of input size and depth of the network are not compatible because of the consecutive use of pooling steps.

Since the number of configurations becomes huge, this first study focuses on finding the optimal parameters for the binary format of the dataset. We shall assume that the best configuration for the binary images of the contest will also achieve good results in other scenarios.

Table 1 shows the F_1 accuracy attained for this case. As can be observed, the different parameters may have some impact on the results but not in a very pronounced way. Unless for very specific cases, all of them achieve a performance higher than 97 %.

For the subsequent experiments, we shall select the configuration that reports the best result in the validation set. Therefore, the CNN chosen to solve the staff-line detection problem, hereinafter referred to as *StaffNet*, is that consisting of 2 layers of 32 3×3 filters plus 2×2 max-pooling, followed by a fully-convolutional layer of 64 units. Activations are computed with a *tanh* function. The input patches must be of size 21×21, centered at the pixel to be classified as either *staff* or *symbol*. A graphical representation of StaffNet can be seen in Fig. 5.

4.2 Comparison with other works

This section details the evaluation methodology carried out to test our proposal against the state-of-the-art, namely the participants of the aforementioned contest. Reader can find a detailed description about each participant in the competition report in [27].

Specifically, we run three different experiments. The first one is focused on reproducing the contest assuming binary images as input. Similarly, the second experiment repeats the same but using the grayscale version of the images of the contest. In the last experiment, another dataset that depicts several deformations on the image is considered, in order to test the robustness of the methods.

4.2.1 Experiment with binary images

Table 2 shows the results (average F_1) achieved by the participants of the contest against the proposed network, considering the binary images of the dataset. We also include the use of conventional classifiers, namely Nearest Neighbor (NN), Support Vector Machines (SVM) and Random Forest (RaF), which were considered in a previous work for solving the same task [4].

It is clear to see that StaffNet outperforms other classification methods, and it represents a remarkable leap in the whole accuracy.

LRDE method also achieves a remarkable performance, even outperforming our method in TS1. It might be that the 3D distortions (that appear in TS1) are harder to handle by our network. Nevertheless, the global accuracy of StaffNet is still higher. The rest of the methods generally show a noticeably worse performance.

These results confirm that the classification approach using StaffNet is indeed a extremely accurate strategy that deserves further consideration. Note that, in this case, the accuracy is not so related to the supervised learning paradigm since other classifiers considered are far from the best performances.

4.2.2 Experiment with grayscale images

As mentioned above, staff-lines removal strategies usually consider a binary image as input. This assumption, however, is not advisable in real-case scenarios. Binarization is a complex process for which it is difficult to

		Filters per layer					
		16		32		64	
Depth	Input	relu	tanh	relu	tanh	relu	tanh
1	07x07	97.42	97.75	97.79	97.82	97.67	97.78
	09x09	98.29	98.27	98.08	98.14	98.39	98.34
	11x11	98.21	98.12	98.17	98.21	97.81	98.34
	13x13	96.07	98.00	98.01	98.05	97.32	98.21
	15x15	96.55	96.52	97.36	97.88	98.35	96.49
	17x17	98.18	97.84	96.19	96.06	96.22	96.19
	19x19	97.78	98.28	96.08	96.21	98.04	97.91
	21x21	97.69	97.93	97.62	98.28	96.25	98.26
	23x23	98.36	98.13	95.97	97.40	97.72	98.24
	25x25	96.11	98.03	98.40	98.51	96.02	96.08
	27x27	98.15	97.66	96.11	98.21	97.12	95.91
	29x29	96.09	95.95	95.86	98.10	98.24	95.93
2	11x11	98.36	98.39	98.59	98.54	98.59	98.71
	13x13	98.29	98.11	98.40	98.40	98.56	98.44
	15x15	97.75	98.74	98.68	98.75	98.82	98.82
	17x17	98.41	98.54	98.78	98.37	98.68	98.70
	19x19	98.65	98.11	98.77	98.68	98.59	98.85
	21x21	98.48	98.65	98.75	98.71	98.85	98.35
	23x23	98.39	98.27	98.80	**99.08**	98.61	98.88
	25x25	98.33	98.71	98.67	98.80	98.51	98.90
	27x27	97.38	98.65	98.53	98.69	98.73	98.86
	29x29	98.69	98.78	98.70	98.52	98.55	98.69
3	23x23	98.58	98.58	98.68	98.89	98.38	98.84
	25x25	98.84	98.75	98.87	98.76	98.96	98.78
	27x27	97.96	98.55	98.60	98.54	98.11	98.72
	29x29	98.68	98.69	98.80	98.80	98.93	98.89

Table 1 F_1 (%) obtained on the validation set for the different network parameters evaluated. Value in bold represent the configuration with the highest performance. *Depth* indicates the number of convolutional plus max-pooling blocks of the model.

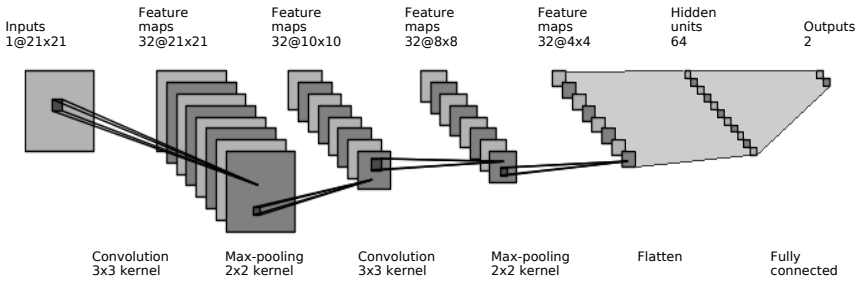


Fig. 5 Graphical representation of the *StaffNet*, consisting of 2 layers of 32 3×3 filters plus 2×2 max-pooling, followed by a fully-convolutional layer of 64 units. Activations are computed with a *tanh* function. The input patches must be of size 21×21 centered at the pixel to be classified as either *staff* or *symbol*.

achieve perfect results, especially when the conditions of the document are not ideal.

The dataset provided in the contest also contains a synthetic grayscale version of the scores. Fortunately, our approach can be easily extended to deal with grayscale images with no further effort. The only thing to change is the training data, which now consists of grayscale patches of the score centered at the pixel to be classified.

Only two of the methods submitted to the contest focused on dealing with grayscale images: LRDE-gray and INESC-gray. Table 3 show the results obtained by these participants, compared to those obtained by our CNN.

Method		TS1	TS2	TS3	Global
Contest participants	TAU	85.72	81.72	82.29	83.01
	NUS	69.85	96.25	67.43	75.24
	NUASI-lin	94.77	94.76	93.81	94.29
	NUASI-skel	94.11	93.67	92.78	93.34
	LRDE-bin	**97.73**	96.86	96.98	97.14
	INESC-bin	89.29	97.72	88.52	91.01
	Baseline	87.01	96.91	89.90	90.93
Conventional classifiers	NN	91.07	96.06	90.58	92.07
	SVM	94.10	98.08	94.00	95.04
	RaF	93.89	97.78	93.39	94.61
StaffNet		97.67	**98.85**	**97.49**	**97.87**

Table 2 F_1 (%) comparison among the participants in the *ICDAR / GREC 2013* staff removal contest, the classifiers of a previous work on staff-line removal, and our method based on CNN (StaffNet) for the binary format of the images. Values in bold represent the best average accuracy in each set.

Method	TS1	TS2	TS3	Global
LRDE-gray	92.16	79.47	81.53	82.85
INESC-gray	38.50	52.11	38.87	42.09
NN	89.65	85.48	85.36	86.46
SVM	92.56	88.84	89.78	90.24
RaF	92.14	86.51	86.50	87.91
StaffNet	**98.91**	**99.29**	**98.64**	**98.87**

Table 3 F_1 (%) comparison among the participants in the *ICDAR / GREC 2013* staff removal contest, the classifiers of a previous work on staff-line removal, and our method based on CNN (StaffNet) for the grayscale format of the images. Values in bold represent the best average accuracy in each set.

It is clear that the performance of the participants decrease remarkably, especially with the INESC method. LRDE is able to maintain a fair accuracy in TS1, but its performance is much worse in TS2 and TS3. Furthermore, our method does improve its recognition by feeding the network with grayscale images. This seems to be related to the supervised learning paradigm since the other classifiers also boost their performance (yet to a lesser extent).

4.2.3 Experiment with distortions

The last experiment focuses on verifying the adaptability of the model to different distortions. It is obvious that, in this case, data-driven strategies have advantages because they are presented with information of the specific domain on which they are going to be applied. It should be noted, however, that traditionally the staff-line removal task have not been been taking into account the great heterogeneity that can be found in musical documents, thus leading to solutions that are not generalizable. We want to demonstrate precisely that it is more profitable to use approaches based on supervised learning, which are more adaptable to other domains by just modifying the training set.

For this experiment we use the MUSCIMA corpus [9], a dataset initially intended for writer identification in musical scores. Fortunately, since this task is hampered by staff lines, the dataset is presented with both original and non-staff examples, allowing us to have a ground-truth readily available. This dataset is composed of 50 writers, each of whom wrote 20 identical scores, and to which they subsequently applied several distortions. Since the writer recognition is not interesting for the task of staff-line detection and removal, we are using the dataset with only the first writer. Therefore, we have 20 images for each of the following distortions:

- Ideal: no distortion applied to the samples.
- Curvature: staves are curved along the x-axis.
- Interrupted: there might be gaps in the staff line segments.
- Kanungo: the degradation model proposed by Kanungo et al. [14] is applied to the whole score.
- Rotated: the score is rotated with respect to the x-axis.
- Thickness (1): the thickness of the staff lines is uniformly increased.
- Thickness (2): the thickness of the staff lines is not regular but depicts different values along the segments.

– Typeset: it mimics a printing mechanism in which little portions of the staff are printed independently.
– Speckles: spurious points are added randomly to the image.
– Y-variation: staff lines are not straight but are placed at different heights.

More details about the distortions and how they were created can be found in the work of Fornes et al. [9].

We ran a new series of new experiments considering only the methods that better behave in the contests, namely INESC and LRDE, as well as SVM as representative of other supervised learning method. We then compare their results against those obtained by StaffNet.

Table 4 reports the performance of the evaluated methods in the new experiment. Some images were rejected by LRDE, that is, nothing is detected (cells marked with a dash). To calculate its average, we have assumed the most beneficial option for LRDE, which is that these cells are not taken into account.

	StaffNet	INESC	LRDE	SVM
Ideal	**99.25**	97.87	97.59	98.51
Curvature	**99.12**	97.43	—	97.97
Interrupted	**99.81**	85.83	—	97.82
Kanungo	**98.57**	96.29	95.25	96.63
Rotated	**99.26**	96.89	—	98.22
Thickness (1)	**98.52**	95.40	—	95.63
Thickness (2)	**97.96**	97.44	96.36	95.85
Typeset	**99.29**	97.72	95.18	98.29
Speckles	**99.16**	97.60	94.00	97.38
Y-variation	**99.10**	97.92	71.06	95.75
Global	**99.00**	96.04	91.57	97.21

Table 4 F_1 (%) comparison among our method, LRDE, INESC and SVM methods for the different distortions of the MUSCIMA dataset. Values in bold represent the best average accuracy in each set. A dash mark (—) is used when the method in the column rejects the dataset of the row (no pixel is categorized in any of its images).

These results reflect the same trend depicted in previous experiments. StaffNet represents a competitive advantage for the problem of staff-line removal, including the case of dealing with different conditions. It can be observed that LRDE, a method that produced excellent results for some data, is seriously harmed by distortions of the input images. INESC, however, is still able to maintain a good performance, yet far from the best performance. SVM can obtain good results as it is also based on learning but, in any case, its results systematically below those obtained by StaffNet.

5 Discussion

Several experiments were presented in the previous section, with the objective of determining if our approach provides a significant improvement with respect to previous methods for the staff-line removal task. Table 5 shows a summary of such experiments.

Experiment	StaffNet	INESC	LRDE	SVM
ICDAR 2013 (bin)	**97.87**	91.01	97.14	95.04
ICDAR 2013 (gr)	**98.87**	42.09	82.85	90.24
MUSCIMA	**99.00**	96.04	91.57	97.21

Table 5 Summary of the F_1 (%) obtained in the evaluated experiments considering StaffNet, LRDE, INESC and SVM methods.

As it can be appreciated, the main problem of the classical methods — represented in this case by INESC and LRDE — is that they depict an irregular behavior, which varies depending on the features of the specific corpus. That is why it is essential that the problem of staff-line removal is approached with supervised learning algorithms.

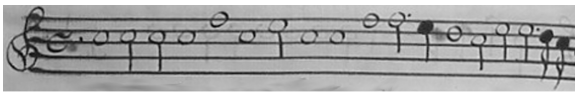
All the experiments carried out dealt with staves of a similar notation (modern), and yet the fact of changing certain characteristics of the documents has already led to unexpected performances in the classical methods. If one is to consider different types of notation and styles, this phenomenon could be even more serious.

Using heuristic methods may be a good idea if the intention is to solve the problem for a particular archive. However, if the intention is to develop methods that can be generalizable, the supervised scenario is the most feasible solution at the moment. Table 5 suggests, however, that it is not enough to consider this paradigm but the most appropriate techniques must be used. As this regard, we have shown that StaffNet leads to a competitive advantage with respect to other supervised learning schemes like SVM, RaF or NN.

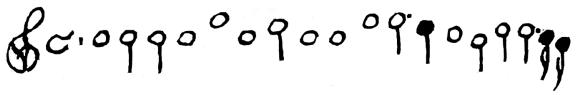
As an illustrative example of these goodnesses, let us consider a new type of musical document. To train StaffNet (and SVM), we are going to consider the ground-truth depicted in Fig. 6. Specifically, this image is of size 1000×250, which contains enough samples to train a classifier from scratch.

Then, a staff-line removal process is applied over the piece of manuscript shown in Fig. 7(a), which depicts Early notation from old music manuscripts. Finally, Fig. 7(b), 7(c) and 7(d) show the performed staff-lines removal in the considered input image by StaffNet, INESC and SVM methods, respectively (LRDE is not included because it rejects the image).

This example illustrates that not only the proposed approach can actually work better for different document types, but that this improvement is important. This can be seen, for example, in the case of symbols broken by other methods, or remaining noise that can be easily confused by musical symbols like the dot.

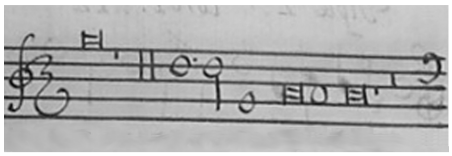


(a) Piece of manuscript of Early notation

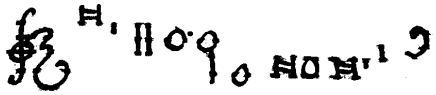


(b) Manually generated ground-truth

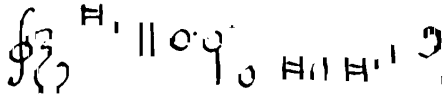
Fig. 6 Qualitative assessment of the staff-line removal methods on Early notation from old music manuscripts.



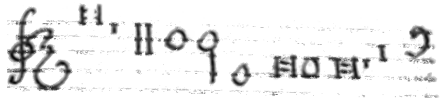
(a) Piece of manuscript of Early notation



(b) Staff-line removal with StaffNet



(c) Staff-line removal with INESC



(d) Staff-line removal with SVM

Fig. 7 Qualitative assessment of the staff-line removal methods on Early notation from old music manuscripts.

Once the goodness of the approach has been discussed, it is important to mention that its obvious drawback is that it requires an appropriate labeled corpus. Yet to this respect, each pixel of an image is a sample of the training set. Therefore, the manual labeling of just a relatively small piece of a score may represent a training set of enough size, as has been demonstrated in the previous example. Furthermore, once a new type of document is to be processed, it is usually cheaper to manually create a labeled corpus than developing a complete new strategy for staff-line removal. In this sense, all experiments have been performed by training the network with data of the same type of document that is eventually presented at the test time. Otherwise, the performance of the model may vary and it cannot be expected that it achieves the performance shown above, unless with scores of similar characteristics. This interesting question will be discussed in future work.

6 Conclusions

This work develops a novel approach to face the music staff removal task, which aims at removing the staff lines from an image of a music score while maintaining the symbol information. This step represents a key stage in most OMR systems. In the literature, staff removal is usually approached by means of image processing techniques based on the intrinsics of music scores. In contrast, we propose to model the problem as an image classification task that can be solved using a CNN. In this context, each foreground pixel is labeled as either *staff* or *symbol* using a square neighborhood as features. Then, the network can be trained using pairs of scores with and without staff lines.

According to our experiments, the approach has demonstrated to be suitable for this task, since the proposed CNN surpassed most of the traditional methods even without applying a post-processing stage (for instance, isolated pixels classified as staff could have been removed). Our method depicted an extremely competitive performance, achieving the highest accuracy in most of the test sets considered (and almost the same accuracy in the other one), and the highest accuracy on average. In addition, we also discussed several advantages of this approach for which conventional methods are not applicable, such as its adaptability to any type of music score.

As a future work, it would be interesting to consider data augmentation techniques to generate more training data depicting different conditions. This is expected to make the neural network generalize better [29]. How-

ever, this is not a trivial matter as, for the task at hand, the augmentation should be analyzed carefully since distortions in music documents are far beyond simple scaling or rotation.

On the other hand, we are interested in the use of fine-tuning strategies to adapt a trained network to process different styles or notations using only a few labeled samples [12].

References

1. Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.
2. Donald Byrd and Jakob Grue Simonsen. Towards a standard testbed for optical music recognition: Definitions, metrics, and page images. *Journal of New Music Research*, 44(3):169–195, 2015.
3. Jorge Calvo-Zaragoza, Isabel Barbancho, Lorenzo J. Tardón, and Ana M. Barbancho. Avoiding staff removal stage in optical music recognition: application to scores written in white mensural notation. *Pattern Anal. Appl.*, 18(4):933–943, 2015.
4. Jorge Calvo-Zaragoza, Luisa Micó, and Jose Oncina. Music staff removal with supervised pixel classification. *International Journal on Document Analysis and Recognition (IJDAR)*, pages 1–9, 2016.
5. Dan Ciresan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3642–3649. IEEE, 2012.
6. Christoph Dalitz, Michael Droettboom, Bastian Pranzas, and Ichiro Fujinaga. A Comparative Study of Staff Removal Algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(5):753–766, 2008.
7. J. Dos Santos Cardoso, A. Capela, A. Rebelo, C. Guedes, and J. Pinto da Costa. Staff Detection with Stable Paths. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(6):1134–1139, June 2009.
8. A. Dutta, U. Pal, A. Fornes, and J. Llados. An Efficient Staff Removal Approach from Printed Musical Documents. In *2010 20th International Conference on Pattern Recognition (ICPR)*, pages 1965–1968, Aug 2010.
9. Alicia Fornés, Anjan Dutta, Albert Gordo, and Josep Lladós. CVC-MUSCIMA: a ground truth of handwritten music score images for writer identification and staff removal. *International Journal on Document Analysis and Recognition*, 15(3):243–251, 2012.
10. Alicia Fornés, Van Cuong Kieu, Muriel Visani, Nicholas Journet, and Anjan Dutta. The ICDAR/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*, pages 207–220, 2013.
11. Thierry Géraud. A Morphological Method for Music Score Staff Removal. In *Proceedings of the 21st International Conference on Image Processing (ICIP)*, pages 2599–2603, Paris, France, 2014.
12. Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
13. Nina S. T. Hirata. Multilevel Training of Binary Morphological Operators. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(4):707–720, 2009.
14. Tapas Kanungo, Robert M Haralick, and Ihsin Phillips. Global and local document degradation models. In *Document Analysis and Recognition, 1993., Proceedings of the Second International Conference on*, pages 730–734. IEEE, 1993.
15. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
16. Igor dos Santos Montagner, Roberto Hirata, and Nina S.T. Hirata. A Machine Learning Based Method for Staff Removal. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 3162–3167, Aug 2014.
17. Weronika Piatkowska, Leszek Nowak, Marcin Pawlowski, and Maciej Ogorzalek. Stafflines Pattern Detection Using the Swarm Intelligence Algorithm. In Leonard Bolc, Ryszard Tadeusiewicz, Leszek J. Chmielewski, and Konrad Wojciechowski, editors, *Computer Vision and Graphics*, volume 7594 of *Lecture Notes in Computer Science*, pages 557–564. Springer Berlin Heidelberg, 2012.
18. Carolina Ramirez and Jun Ohya. Automatic recognition of square notation symbols in western plainchant manuscripts. *Journal of New Music Research*, 43(4):390–399, 2014.
19. A. Rebelo and J.S. Cardoso. Staff Line Detection and Removal in the Grayscale Domain. In *2013 12th International Conference on Document Analysis and Recognition (ICDAR)*, pages 57–61, Aug 2013.
20. Ana Rebelo, G Capela, and Jaime S Cardoso. Optical recognition of music symbols. *International Journal on Document Analysis and Recognition (IJDAR)*, 13(1):19–31, 2010.
21. Ana Rebelo, Ichiro Fujinaga, Filipe Paszkiewicz, André R. S. Marçal, Carlos Guedes, and Jaime S. Cardoso. Optical music recognition: state-of-the-art and open issues. *IJMIR*, 1(3):173–190, 2012.
22. Florence Rossant and Isabelle Bloch. Robust and adaptive OMR system including fuzzy modeling, fusion of musical rules, and possible error detection. *EURASIP Journal on Applied Signal Processing*, 2007(1):160–160, 2007.
23. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.
24. Bolan Su, Shijian Lu, U. Pal, and C.L. Tan. An Effective Staff Detection and Removal Technique for Musical Documents. In *2012 10th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 160–164, March 2012.
25. Lorenzo J. Tardón, Simone Sammartino, Isabel Barbancho, Verónica Gómez, and Antonio Oliver. Optical Music Recognition for Scores Written in White Mensural Notation. *EURASIP J. Image and Video Processing*, 2009, 2009.
26. Rainer Typke, Frans Wiering, and Remco C. Veltkamp. A survey of music information retrieval systems. In *ISMIR 2005, 6th International Conference on Music Information Retrieval, London, UK, 11-15 September 2005, Proceedings*, pages 153–160, 2005.
27. M. Visaniy, V.C. Kieu, A. Fornes, and N. Journet. ICDAR 2013 Music Scores Competition: Staff Removal. In *2013 12th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1407–1411, Aug 2013.

28. Matthew D. Zeiler. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701, 2012.
29. Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer vision–ECCV 2014*, pages 818–833. Springer, 2014.