CVC-MUSCIMA: A Database of Handwritten Music Score Images for Writer Identification and Staff Removal

Alicia Fornés · Anjan Dutta · Albert Gordo · Josep Lladós

1 Introduction

The analysis of music scores has been an active research field in the last decades. However, there are no public available databases of handwritten music scores for the research community. In this paper we present the CVC-MUSCIMA database of handwritten music score images. It consists of 1,000 music sheets written by 50 different musicians. The dataset has been especially designed for writer identification and staff removal tasks. We describe the evaluation metrics, partitioning and baseline results for easing the comparison between different approaches.

Keywords Music Scores · Handwritten Documents · Writer Identification · Staff Removal · Performance Evaluation · Graphics Recognition · Ground-truths

Abstract The analysis of music scores has been an active research field in the last decades. However, there are no public available databases of handwritten music scores for the research community. In this paper we present the CVC-MUSCIMA database of handwritten music score images. It consists of 1,000 music sheets written by 50 different musicians. The dataset has been especially designed for writer identification and staff removal tasks. We describe the evaluation metrics, partitioning and baseline results for easing the comparison between different approaches.

Keywords Music Scores · Handwritten Documents · Writer Identification · Staff Removal · Performance Evaluation · Graphics Recognition · Ground-truths

1 Introduction

The analysis of music scores has been a very active research topic in the last decades [11, 15, 18, 20]. Traditionally, the main focus of interest within the research community has been the transcription of printed music scores. Optical Music Recognition (OMR) [1] consists in the understanding of information from digitized music scores and its conversion into a machine readable format. It allows a wide variety of applications such as the edition of scores never edited, renewal of old scores, conversion of scores into braille, production of audio files, etc. Among the required stages of an Optical Music Recognition system, an special emphasis has been put in the staff removal algorithms [3, 8, 19], since a good detection and removal of the staff lines will allow the correct isolation and segmentation of the musical symbols, and consequently, will ease the correct detection, recognition and classification of the music symbols.

Lately, there has been a growing interest in the analysis of handwritten music scores [16, 17]. In this context, the focus of interest is two-fold: the recognition of handwritten music scores, and the identification (or verification) of the authorship of a music score. Concerning writer identification, musicologists do not only perform a musicological analysis of the composition (melody, harmony, rhythm, etc.), but also analyse the handwriting style of the manuscript. In this sense, writer identification can be performed by analyzing the shape of the hand-drawn music symbols (e.g. music notes, clefs, accidentals, rests, etc.), because it has been shown that the author’s handwriting style that characterizes a piece of text is also present in a graphic document. Nevertheless, musicologists must work very hard to identify the writer of a music score, especially when there is a large amount of writers to compare with. Recently, several writer identification approaches have been developed for helping musicologists in such a time consuming task. These approaches are based in many different methodologies, such as Self Organizing Maps [14], Bag of Features [10], knowledge-based approaches [2, 9], or even systems which adapt some writer identification approaches for text documents to music scores [7].

Contrary to printed music scores databases [3], there are no public databases of handwritten music scores available for the research community. For this reason, there is a need of a public database and ground-truth for validating the different methodologies developed in this research field. With this motivation, in this pa-
per we present the CVC-MUSCIMA \(^1\) \(^2\) database: a database of handwritten music score images. This dataset consists of 1,000 music sheets written by 50 different musicians, and has been especially designed for writer identification and staff removal tasks. We include the evaluation metrics, partitions (data subsets) and baseline results for comparison purposes. We believe that the presented ground-truth will serve as a basis for research in handwritten music analysis.

The rest of the paper is organized as follows. Section 2 describes the dataset and the staff distortions applied. Section 3 presents the evaluation partitions, metrics and baseline results for comparison purposes. Finally, concluding remarks are described in Section 4.

**2 Dataset**

The dataset consists of 20 music pages of different compositions transcribed by 50 writers, yielding a total of 1,000 music sheet pages. All the 50 writers are adult musicians in order to ensure that they have their own characteristic handwriting music style. Each writer has been asked to transcribe exactly the same 20 music pages, using the same pen and the same kind of music paper (with printed staff lines). The set of the 20 selected music sheets contains monophonic and polyphonic music, and it consists of music scores for solo instruments (e.g., violin, flute, violoncello or piano) and music scores for choir and orchestra. It must be noted that the music scores only contain the handwriting text considered as part of the music notation theory (such as dynamics and tempo notation), and for this reason, music scores for choir do not contain lyrics.

Furthermore, for staff removal tasks, each music page has been distorted using different transformation techniques (please refer to Section 2.2 for details), which, together with the originals, yields a grand total of 12,000 base images.

Next, we describe the data acquisition, the generated deformations and the different ground-truths and data formats.

**2.1 Acquisition and Preprocessing**

Documents were scanned using an flatbed Epson GT-3000 scanner set at 300 dpi and 24 bpp, as colour cues were used in the original templates to ease the elaboration of the staff ground-truth. Later, the images were converted to 8 bit gray scale. Care was put into obtaining a good orientation during the scanning stage, and absolutely no digital skew correction was applied once the pages were scanned.

The staff lines were initially removed using color cues. Afterwards, they were binarized and manually checked for correcting errors, specially when some segments of the staff lines were manually added by the writer (see an example in Fig. 1). Thus, from the gray scale images, we generated the binarized images, the images with only the music symbols (without staff lines), and finally, the images with only the staff lines. Next, we describe the distortions applied to the music scores for staff removal.

![Fig. 1: Example of a section of a music score with some segments of hand-drawn staff lines](image)

**2.2 Staff Distortions**

To test the robustness of different staff removal algorithms, we have applied a set of distortion models to our music score images. These distortion models are inspired by the work of Dalitz et al. [3] for testing the performance of staff removal algorithms in printed music scores. In [3] the authors describe nine different types of deformations for simulating their dataset with real world situation: Degradation with Kanungo noise, Rotation, Curvature, Staffline interruption, Typeset emulation, Staffline y-variation, Staffline thickness ratio, Staffline thickness variation and White speckles.

In order to obtain the same effect, the deformation is simultaneously applied to the original and the ground-truth staff images, which correspond to binary images with only the staff lines. A brief description of the individual deformation models is given next:

\(^1\) CVC-MUSCIMA stands for Computer Vision Center - Music Score Images

\(^2\) http://www.dag.cvc.uab.es/cvcmuscima
– Kanungo noise. Kanungo et al [13] have proposed a noise model to create local distortions introduced during scanning. The model mainly affects to the contour pixels and has little effect on the inner pixels (see Fig. 2b).

– Rotation. The distortion rotation (see Fig. 2c) consists in rotating the entire staff image by the specified parameter angle.

– Curvature. The curvature is performed by applying a half sinusoidal wave over the entire staffwidth. The strength of the curvature is regulated by a parameter which is a ratio of the amplitude to the staffwidth (see Fig. 2d).

– Staffline interruptions. The staffline interruptions consist in generating random interruptions with random size in the stafflines. This model mainly affects to the staffline pixels, and simulates the scores that are written on already degraded stafflines (see Fig. 2e).

– Typeset emulation. This particular defect is intended to imitate the sixteenth century prints that are set by lead types. Consequently, they have staffline interruptions between symbols and also a random vertical shift of each vertical staff slice containing a symbol (see Fig. 2f).

– Staffline y-variation and Staffline thickness variation. These kind of defects are created by generating a Markov chain describing the evolution of the y-position or the thickness from left to right. This is done since, generally the y-position and the staff thickness values for a particular x-position depend on its previous x-position (Fig. 2g, 2h, 2i, and 2j show some examples of these deformations with different parameters).

– Staffline thickness ratio. This defect only affects to the whole staffline thickness of the music score, which consists in generating stafflines of different thickness (see Fig. 2k).

– White speckles. This degradation model is used to generate white noise within the staff pixels and musical symbols (see Fig. 2l).

Table 1 describes the parameters of the respective models. Dalitz et al. [3] have developed the MusicStaves toolkit 3, which is available for reproducing the experiments in other datasets. However, these available algorithms for distorting the staff lines have an important drawback: they require computer generated perfect artificial images, which means perfect horizontal staff lines, equidistant, and also with the same thickness. Since our dataset contains printed and handwriting.


Fig. 2: Staff deformation and their corresponding parameters. (a) Ideal image. (b) Kanungo \((\eta, \alpha_0, \alpha, \beta_0, \beta, k) = (0, 1, 1, 1, 2)\). (c) Rotation \((\theta) = (12.5^\circ)\). (d) Curvature \((a, p) = (0.05, 1.0)\). (e) Staffline interruptions \((a, n, p) = (0.5, 3, 0.5)\). (f) Typeset emulation \((n, p, ns) = (1, 0.5, 10)\). (g)-(h) Staffline y-variation \((n, c) = (5, 0.6)\) and \((n, c) = (5, 0.93)\). (i) Staffline thickness ratio \((r) = (1.0)\). (j)-(k) Staffline thickness variation \((n, c) = (6, 0.5)\) and \((n, c) = (6, 0.93)\). (l) White speckles \((p, n, k) = (0.025, 10, 2)\).
Table 1: Image deformation and corresponding parameters. For more information about the parameters and the generation of each distortion, refer to [3].

<table>
<thead>
<tr>
<th>Deformation</th>
<th>Parameters description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kanungo noise</td>
<td>Each foreground pixel is flipped with probability ( q e^{-\eta d^{\alpha} + \eta} ) (( d ) is the distance to the closest background pixel); each background pixel is flipped with probability ( \beta e^{-\beta d^{\alpha} + \eta} ) (( d ) is the distance to the closest foreground pixel).</td>
</tr>
</tbody>
</table>

| Rotation (\( \theta \)) | \( \theta \) is the rotation angle to be applied. |
| Curvature (\( a, p \)) | \( a \) is the amplitude of the sine wave divided by the staffwidth, \( p \) is the number of times a half sine wave should appear in the entire staff width. |
| Stalline interruption (\( \alpha, n, p \)) | \( \alpha \) is the probability for each pixel to be the center of an interruption, \( n \) and \( p \) are the parameters for the binomial distribution of size interruption. |
| Typeset emulation (\( n, p, ns \)) | \( n \) and \( p \) are the parameters for the binomial distribution deciding the horizontal gaps, \( ns \) is the parameter for another binomial distribution deciding the y gap where the value of the other parameter is always 0.5. |
| Stalline y-variation (\( n, c \)) | \( n \) and \( p = 0.5 \) are the parameters of the binomial distribution deciding the stationary distribution of Markov chain, \( c \) is an inertia factor allowing the smooth transition. |
| Stalline thickness (\( r \)) | \( r \) is the ratio of the staffline height to staffspace height. |
| Stalline thickness variation (\( n, c \)) | \( n \) and \( p = 0.5 \) are the parameters of the binomial distribution deciding the stationary distribution of Markov chain, \( c \) is an inertia factor allowing the smooth transition. |
| White speckles (\( p, n, k \)) | \( p \) is the parameter for speckle frequency, \( n \) is the size of the speckle and \( k \) is the size of the structuring element used for closing operation. |

Table 2: Image flavours designed for writer identification and staff removal tasks. Recommended images for each task in bold.

<table>
<thead>
<tr>
<th>Task</th>
<th>Images provided</th>
</tr>
</thead>
<tbody>
<tr>
<td>Writer</td>
<td>1,000 original undistorted grey scale images</td>
</tr>
<tr>
<td>Ident.</td>
<td>1,000 binary staffless images</td>
</tr>
<tr>
<td>Staff</td>
<td>12,000 binary images with staff lines</td>
</tr>
<tr>
<td>Removal</td>
<td>12,000 binary images of only staff lines</td>
</tr>
<tr>
<td></td>
<td>12,000 binary staffless images</td>
</tr>
</tbody>
</table>

3 Ground-truth

In this section we describe the images, evaluation partitions (subsets), evaluation metrics and some baseline results. Thus, they will serve as a benchmark scenario for a fair comparison between different approaches.

Concerning the baseline results, it must be said that since the main contribution of this work is the framework for performance evaluation, we include some baseline results just for reference purposes.

3.1 Images Description

All the images of the dataset are presented in PNG format. Each document of the dataset (1,000 original images plus the 11,000 distorted images) is labelled with its writer identification code and presented in different image flavours:

- Original grey scale image (only for the original 1,000 images).
- Binary image (with staff lines).
- Binary staffless image (only music symbols).
- Binary staff lines image (no music symbols).

Although all this information is available for all tasks, we encourage the use of certain image flavours for different tasks. The staffless images are particularly useful for writer identification: since most writer identification methods remove the staff lines in the preprocessing stage, this eases the publication of results which are not dependant on the performance of the particular staff removal technique applied. See an example in Fig. 3. Similarly, for the staff removal tasks, staff lines images without music symbols (see Fig. 4) may be useful, not only for the evaluation of the method but also for training purposes. Table 2 summarizes the provided images and these recommendations.
3.2 Evaluation Partitions for Writer Identification

For training and evaluation purposes, we devised two sets of ten partitions, which were especially designed for the evaluation of writer identification tasks:

**Set A (or Dependent).** In the first set of partitions, the training pieces of a given partition are the same for each writer, and thus, none of the pieces of the test set have been seen during the training stage. As an illustrative example, if the 1st music page of one writer is in the test set of a given partition, all the 1st music pages of the remaining writers will be also in the test set of that partition. Since train and test pieces are the same for each writer, is not possible to confuse a testing piece with the same piece written by a different author (since those will not be in the training set), and because of this we will refer to this set as dependent.

**Set B (or Independent).** In the second set of partitions this constraint is not satisfied, and pieces that appear in the training set of one author will appear in the test set of a different one (for example, the 1st music page will appear in the train set of one author and in the test set of another). We will refer to this set as independent. These partitions are particularly devised to attest that we are indeed performing writer identification instead of rhythm classification, since a piece with exactly the same rhythm but from a different writer will appear in the training set. Therefore, a high writer identification rate in this group of partitions will show that the system is classifying according to the handwriting style and not being affected by the kind of music notes and symbols appearing in the music sheet.

In each partition, 50% of the documents of each writer belong to the training set and the other 50% belong to the test set. Furthermore, effort has been put in guaranteeing that each piece appears approximately 50% of the time in training and 50% in test. The exact partitions can be found in the dataset pack.

3.3 Evaluation Metrics and Baseline Results for Writer Identification

In this subsection, we will describe the evaluation metrics and, as an illustrative example, some baseline results for writer identification purposes.


**Metrics.** Some writer identification systems are evaluated considering two options: if the image has been correctly classified taking into account the n-first authors, or only the first writer. In our scenario, we will treat it as a binary problem, in which a music score is correctly classified only if the first nearest writer corresponds to the ground-truthed one.

**Method.** As we have stated, it is out of the scope of this work to make a comparison of different writer identification methods in the literature. However, and only for reference purposes, we provide baseline results using a recent writer identification method for musical scores.

In the Bag-of-Notes approach described in [10], features are represented using the Blurred Shape Model descriptor [5]. Then, a probabilistic codebook is built using a Gaussian Mixture Model and soft assignment is used to represent the musical scores. Finally, they are classified using a SVM with an RBF kernel.

Our implementation of the Bag-of-Notes slightly differs from the one presented in [10]. First, the work presented in [10] reports results using both unsupervised and supervised clustering, which means learning the codebook with all the authors vs. learning a different codebook for each author and then merging them. We did not observe a significant difference in the results, so, for simplicity, we only report results with unsupervised clustering. Second, instead of using an RBF kernel, we used a linear SVM, since we did not observe any significant improvement of the RBF kernel over the linear SVM. Moreover, using a linear kernel allows us to use solvers optimized for linear problems such as LIBLINEAR [6], which makes use of the cutting-plane algorithm and drastically improves the training speed of the SVM. To set the C trade-off cost of the SVM classifier, we used the same heuristic used by the SVM\textsuperscript{light} [12] suite. Given a set of N training vectors $X = \{x_1, x_2, \ldots, x_N\}$, we set $C$ as follows:

$$C = 1/k^2, \quad k = \frac{1}{N} \sum_{i=1}^{N} x_i x'_i$$

(1)

This heuristic gave excellent classifications results, in fact superior to those obtained by manually setting the parameter.

**Results.** Table 3 reports mean classification accuracy and standard deviation as a function of the number of Gaussians for the two sets of partitions (please cf. Section 3.2 for details on these partitions). Note that the accuracy results on both sets are quite similar, with a slight advantage for the second set; the higher accuracy and smaller standard deviation are probably caused because this set contains more variety in the training data. The fact that both sets obtain very similar results suggests that the Bag-of-Notes method is indeed performing writer identification and not rhythm identification, as would be the case if the dependent set obtained significantly better results than the independent set.

### Table 3: Mean classification accuracy (in %) and standard deviation as a function of the number of Gaussians.

<table>
<thead>
<tr>
<th>N. of Gaussians</th>
<th>Set A (dep.)</th>
<th>Set B (indep.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>31.60 ± 4.28</td>
<td>31.82 ± 2.06</td>
</tr>
<tr>
<td>32</td>
<td>43.18 ± 3.73</td>
<td>45.36 ± 3.36</td>
</tr>
<tr>
<td>64</td>
<td>57.08 ± 4.15</td>
<td>59.20 ± 2.83</td>
</tr>
<tr>
<td>128</td>
<td>73.02 ± 3.83</td>
<td>75.00 ± 2.06</td>
</tr>
<tr>
<td>256</td>
<td>84.72 ± 3.20</td>
<td>86.12 ± 1.35</td>
</tr>
</tbody>
</table>

3.4 Evaluation Metrics and Baseline Results for Staff Removal

In this subsection, we will describe the evaluation metrics and some baseline results for staff removal purposes.

**Metrics.** According to [3], we have chosen the pixel based evaluation metric to get the quantitative measurement of the performance of our algorithm. Here the staff detection or staff removal problem is considered as a two-class classification problem at the pixel level. For each of the images we compute the number of true positive pixels $tp$ (pixels correctly classified as staff lines), false positive pixels $fp$ (pixels wrongly classified as staff lines) and false negative $fn$ (pixels wrongly classified as non-staff lines) by overlapping with the corresponding ground truth images. The precision and recall measures of the classification are computed as:

$$\text{Precision} = P = \frac{tp}{tp + fp}, \quad \text{Recall} = R = \frac{tp}{tp + fn}$$

(2)

The third metric error rate $E$ is computed as (\# means ”number of”, $sp$ means ”staff pixels”):

$$E = \frac{\#\text{misclassified } sp + \#\text{misclassified non } sp}{\#\text{all } sp + \#\text{all non } sp}$$

(3)

**Method.** For the sake of illustration, we have chosen one of our staff removal algorithms as the baseline results. The approach proposed in [4] is based on the criteria of neighboring staff components. It considers a staffline segment as a horizontal linkage of vertical black
runs with uniform height, and then it uses the neighbouring properties of a staffline segment to discard the false segments.

Results. Table 4 shows the results of the staff removal algorithm using the proposed evaluation metrics and applied to the 12,000 distorted images. It must be noted that it does not obtain the best results in all cases with respect to the three evaluation metrics, showing that there is still room for research in this field. It should also be noted that these results are over the whole dataset and not the testing set, since this method do not require any training step.

4 Conclusions

In this paper we have described the CVC-MUSCIMA database and ground-truth, which has been especially designed for writer identification and staff removal tasks. The database can serve as a basis for research in music analysis. Moreover, we have described the evaluation metrics, partitions and baseline results in order to ease the comparison between the different approaches that may be developed.

The database and ground-truth is considered complete at the current stage. However, further work could be focused on labelling each music note and symbol of the music score images for Optical Music Recognition purposes.

Acknowledgements. We would like to thank all the musicians who contributed to the database presented in this paper. We would also specially thank to Joan Casals from the Universitat Autònoma de Barcelona for contacting with musicians, and collecting the music sheets. We would also like to thank Dr. Christoph Dalitz for providing the code which generates the staff distortions. This work has been partially supported by the Spanish projects TIN2008-04998, TIN2009-14633-C03-03, and CONSOLIDER-INGENIO 2010 (CSD2007-00018).

References


