The ICDAR 2013 Music Scores Competition: Staff Removal

Muriel Visani[†], V.C Kieu[*†], Alicia Fornés [‡], and Nicholas Journet[*]
[†]Laboratoire Informatique, Image et Interaction - L3i, University of La Rochelle, La Rochelle, France
[*]Laboratoire Bordelais de Recherche en Informatique - LaBRI, University of Bordeaux I, Bordeaux, France
[‡]Computer Vision Center - Dept. of Computer Science, Universitat Autònoma de Barcelona, Ed.O, 08193, Bellaterra, Spain
Email: muriel.visani@univ-lr.fr, {vkieu, journet}@labri.fr, afornes@cvc.uab.es

***Abstract*—The first competition on music scores that was organized at ICDAR in 2011 awoke the interest of researchers, who participated both at staff removal and writer identification tasks. In this second edition, we focus on the staff removal task and simulate a real case scenario: old music scores. For this purpose, we have generated a new set of images using two kinds of degradations: local noise and 3D distortions. This paper describes the dataset, distortion methods, evaluation metrics, the participant's methods and the obtained results.**

***Keywords*—Competition, Music Scores, Staff Removal.**

I. Introduction

The recognition of music scores has been an active research field for decades [1], [2]. Many researchers have proposed staff removal algorithms in order to improve the recognition of music symbols [3], [4]. However, the staff removal task cannot be considered as a solved problem, especially when dealing with ancient/degraded handwritten music scores.

At ICDAR [5] and GREC 2011, we organized the first edition of the music scores competition. For the staff removal task, we generated distorted images in order to test the robustness of the staff removal algorithms. Thus, we created several sets of distorted images, where each set corresponded to a different kind of distortion (e.g. Kanungo noise, rotation, curvature, staff line interruption, typeset emulation, staff line thickness variation, white speckles, etc.).

After GREC 2011, we extended the staff removal competition [6]. The goal was to simulate a real scenario, in which music scores usually contain more than a single kind of distortion. For this purpose, we generated a new set of images, by combining different distortions at different levels. Not surprisingly, the new results demonstrated that for most methods, the performances were significantly decreased when coping with a combination of distortions.

By organizing a second edition of this competition at ICDAR/GREC 2013, we aimed at fostering the interest of researchers we attracted during the preceding edition. For this second edition, we have generated new images that emulate typical degradations appearing in old and degraded handwritten documents. In this new edition, two types of degradations have been considered: local noise and 3D distortions.

The rest of the paper is organized as follows. First, we describe the original dataset, the degradation models, the training and test sets and the evaluation metrics. Afterwards, we present the participants' methods and analyze the results.

II. The ICDAR/GREC 2013 Music Scores Database

For the music scores competition on staff removal, we generated a semi-synthetic database (presented in section II-C) using the original CVC-MUSCIMA database (see section II-A) and two degradation models we introduced recently (see section II-B).

A. Original CVC-MUSCIMA database

The original CVC-MUSCIMA [1] database [7] consists of handwritten music score images. An extract of this database can be seen in Fig. 1(a). The dataset contains 1,000 music sheets written by 50 different musicians. Each musician has transcribed exactly the same 20 music pages, using the same pen and the same kind of music paper. The 20 selected music sheets contain monophonic and polyphonic music.

B. Degradation Models

1) 3D Degradation Model: Since the geometric distortions such as skews and curvatures generated by 2D distortion models are challenging for detecting staffs, we used them for the ICDAR/GREC 2011 staff removal competition [5], [6]. However, these 2D distortions models cannot reproduce the geometric distortions commonly encountered in real old documents such as dents, small folds, torns.... In this 2013 edition of the ICDAR/GREC staff removal competition, we use the 3D distortion model we introduced in [8]. It can generate more realistic and challenging distortions of the staff lines, making their detection and removal more difficult. This 3D degradation model is based on 3D meshes and texture coordinate generation. The main idea is to wrap any 2D (flat) image of a document on a 3D mesh acquired by scanning a non-flat old document using a 3D scanner. The wrapping functions we use are specifically adapted to document images. In our case, we wrap the original images of the MUSCIMA database on different 3D meshes we previously acquired.

2) Local Noise Model: The local noise model we introduced in [9] is specifically designed to mimic some old documents' defects such as ink splotches and white specks or streaks. These defects might break the connectivity of strokes or, inversely, add a connection between separated strokes. In our case, the application of this kind of degradation to the MUSCIMA database can lead to disconnections of the staff lines, or the addition of dark specks connected to a staff

[1] Available at http://www.cvc.uab.es/cvcmuscima/

line. In the latter case, the dark specks might be confused with musical symbols. More generally, local noise can lead to many degradations which are very challenging for staff removal algorithms.

As detailed in [9], the local noise is generated in three main steps. Firstly, the "seed-points" (*i.e.* the centres of local noise regions) are selected so that they are more likely to appear near the foreground pixels (obtained by binarizing the input grayscale image). Then, we add arbitrary shaped grey-level specks (in our case the shape is an ellipse). The grey-level values of the pixels inside the noise regions are modified so as to obtain realistic looking bright and dark specks.

C. The ICDAR/GREC 2013 Database

The ICDAR/GREC 2013 database consists in 6000 images (4000 for training, 2000 for testing), as detailed hereafter:

1) Training Set: The training set consists in 4000 semi-synthetic images generated from 667 out of the 1000 original images in the CVC-MUSCIMA database. It is split into three subsets corresponding to different degradation types:

- *TrainingSubset1* contains 1000 images generated using the 3D distortion model (*c.f.* sub-section II-B1) and two different meshes (with 500 images per mesh). As it can be seen in Fig. 1, the first mesh contains essentially a perspective distortion due to the scanning of a thick and bound page, while the second mesh has many small curves, folds and concaves.
- *TrainingSubset2* (see Fig. 2) contains 1000 images generated with three different levels of local noise, where a higher level of noise corresponds to an increasing number of seed points and larger noise regions (see section II-B2).
- *TrainingSubset3* (see Fig.3) contains 2000 images from six degradation levels obtained by combining the same three levels of local noise and 3D distortions (with the two 3D meshes used in TrainingSubset1).

For each image in the training set, we provide to the participants of the competition its grey and binary versions and the associated ground-truth, under the form of its binary staff-less version, as illustrated in Fig. 4.

2) Test Set: The test set consists in 2000 semi-synthetic images generated using the 333 original images from the CVC-MUSCIMA database that are not used for the training set.

- *TestSubset1* contains 500 images generated using the 3D distortion degradation model and two meshes (distinct from the ones used in the training set).
- *TestSubset2* contains 500 images generated using the same three levels of local noise as in TrainingSubset2.
- *TestSubset3* contains 1000 images equally distributed between six different levels of degradation obtained by combining the same three levels of local noise and 3D distortions (with the 2 meshes used in TestSubset1).

For each image in the test set, we provide to the participants of the competition its gray and binary versions. The ground-truth associated to the test set, consisting of binary staff-less images, was made public after the contest.

III. Experimental Protocol and Results

The competition was organized in three steps. First, we provided to the participants (see section III-A) the training set and its grount-truth for training their algorithms. Second, we sent them the test set 46 days later. Third, they returned us their outputs as binary staff-less images within 23 days. Further, we compared their outputs to the ground-truth of the test set using the evaluation measures in section III-C. The obtained results are confronted to the results of a baseline method (see section III-B) and analyzed in section III-D.

A. Participants Information

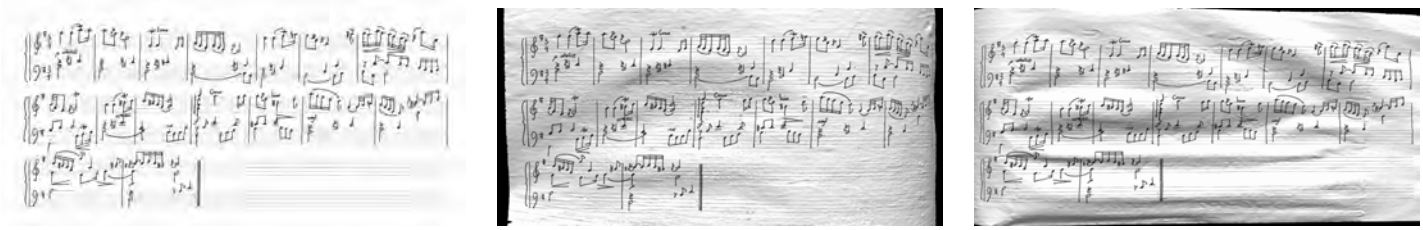
In this section, we briefly describe the methods used by the participants for the ICDAR/GREC2013 competition. Methods 1-3 work on binary images (in that case the participants used the binary versions we provided for the competition), while methods 4-5 can handle both binary and grayscale images.

1) TAU-bin: This method was submitted by Oleg Dobkin from the Tel-Aviv University, Israel. It is based in the Fujinaga's method [10]. First, the *staffline_height* and *staffspace_height* are estimated using vertical scans. Then, the vertical black runs which are longer than the *staffspace_height* are removed. Afterwards, the music page is deskewed, and the staff lines are located using a projection on the y-axis. Finally, the staff lines are removed using masks.

2) NUS-bin: This method was submitted by Bolan Su (National University of Singapore), Umapada Pal (Indian Statistical Institute, Kolkata, India) and Chew-Lim Tan (National University of Singapore). The method, detailed in [11], first estimates the *staffline_height* and *staffspace_height* using the vertical run length histogram. These estimated values are used to predict the lines' direction and fit an approximate staff line curve for each line. Then, the fitted staff line curve is used to identify the exact location of staff lines, and those pixels belonging to these staff lines are removed.

3) NUASi: Christoph Dalitz and Andreas Kitzig, from the Niederrhein University of Applied Sciences (iPattern Institute), Krefeld, Germany, submitted two different methods described shortly hereafter:

- NUASi-bin-lin: This method is described in Section II of [3]. First, the *staffline_height* is estimated as the most frequent black vertical run length. Then, the skeleton of the staff lines is extracted, and all vertical foreground runs shorter than 2* *staffline_height* are removed. The source code is available at *http://music-staves.sourceforge.net/ (class MusicStaves_linetracking)*
- NUASi-bin-skel: This method, detailed in the Section III.D of [3], first splits the skeleton of the staff lines at branching and corner points. Each segment is considered as a staff line segment if it satisfies some heuristic rules. Then, two staff segments are horizontally linked if their extrapolations from the end points with the least square fitted angle are closer than *staffline_height*/2. To check for the false positives, non-staff segments which have the same splitting point with a staff segment are extrapolated by a parametric parabola. If the staff segment is tangent with the



(a) Original CVC-MUSCIMA image

(b) Semi-synthetic image generated using Mesh #1

(c) Semi-synthetic image generated using Mesh #2

Fig. 1. From left to right: original (real) image and two semi-synthetic images from the TrainingSubset1 generated from this image using different 3D meshes

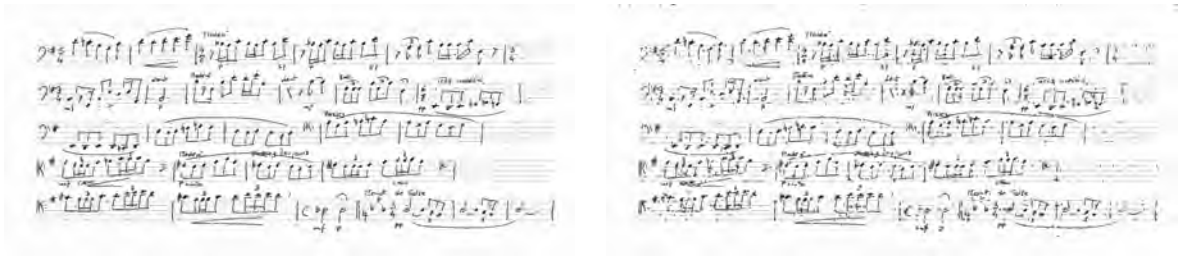


Fig. 2. Two semi-synthetic images from the TrainingSubset2 generated using (respectively from left to right) a low level and a high level of local noise

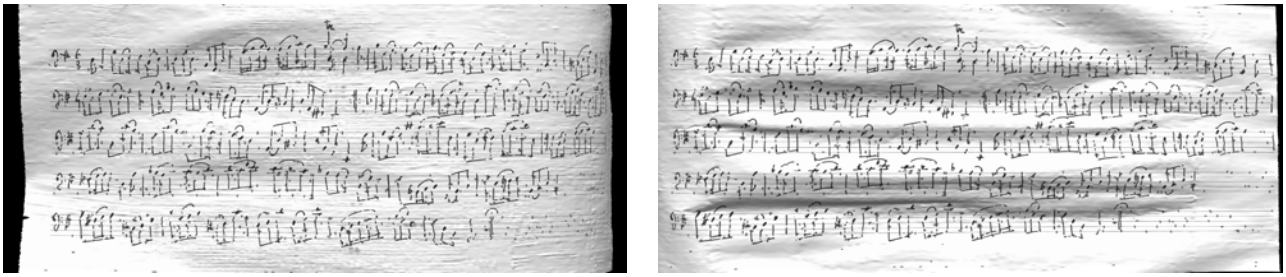


Fig. 3. Two semi-synthetic images from the TrainingSubset3 generated using a high level of local noise and (from left to right) Mesh #1 and Mesh #2

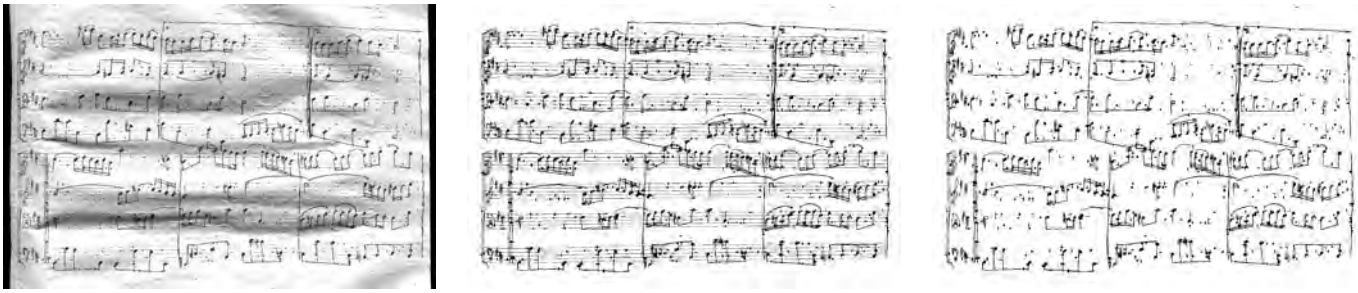


Fig. 4. From left to right: an image from TrainingSubset3, its binary version and its binary staff-less version (ground-truth)

parabola, it is a non-staff segment. Finally, short vertical black runs around the detected staff skeleton are removed. The source code is available at *http://music-staves.sourceforge.net/(class MusicStaves_skeleton)*

4) LRDE: The following two methods (binary and gray versions) were submitted by Thierry Géraud, from the EPITA Research and Development Laboratory (LRDE), Paris, France. More details can be found at *http://www.lrde.epita.fr/cgi-bin/twiki/view/Olena/Icdar2013Score*.

- LRDE-bin: This method relies on mathematical morphological operators. First, a permissive hit-or-miss with a horizontal line pattern as structuring element extracts some horizontal chunks. Second, a horizontal median filter cleans up the result, and a dilation operation is applied using a horizontal neighbourhood in order to enlarge the connected components. A binary mask is obtained thanks to a morphological closing with a rectangular structuring element. Last, a vertical median filter, applied inside the largest components of this mask, removes the staff lines.

- LRDE-gray: After removing the image border, Sauvola's binarization and a dilation using a horizontal

neighbourhood are applied. That image serves as a mask in which a two-level thresholding with hysteresis of the original image is applied. Then, some spurious horizontal parts of staff are erased.

5) INESC: Ana Rebelo and Jaime S. Cardoso (INESC Porto and Universidade do Porto) submitted the following methods (binary and grayscale versions) based on graphs of Strong Staff-Pixels (SSP: pixels with a high probability of belonging to a staff line):

- INESC-bin: First, the *staffline_height* and *staffspace_height* are estimated by the method presented in [12]. Then, all the pixels of the black runs of *staffline_height* pixels followed or preceded by a white run of *staffspace_height* pixels are set as the SSPs. To decide if a SSP belongs to a staff line, the image grid is considered as a graph with pixels as nodes, and arcs connecting neighbouring pixels. Then, SSPs are classified as staff line pixels acording to some heuristic rules. Then, the groups of 5 staff lines are located among the shortest paths by using a global optimization process on the graph (see [4] for details).
- INESC-gray: It applies a sigmoid-based weight function that favors the luminance levels of staff. Then, the image is binarized and the previous method is applied.

B. Baseline

For comparison purposes, we have computed some baseline results using the staff removal method proposed by Dutta *et al.* [13], which is based on the analysis of neighbouring components. Basically, it assumes that a staff line candidate segment is a horizontal linkage of vertical black runs with uniform height. Then, neighbouring properties (e.g. proximity, orientation) are used to validate or discard these segments.

C. Measures Used for Performance Comparison

At the pixel level, the staff removal problem can be seen as a two-class classification task. We compare the participant's images to the test set ground-truth (for each level of degradation), and we compute the number TP of True Positive pixels (pixels correctly classified as staff lines), the number FP of False Positives (wrongly classified as staff lines) and the number FN of False Negatives (wrongly classified as non-staff lines). Then, we compute the F-measure F_M:

$$F_M = 2\times\frac{precision \times recall}{precision + recall} = 2\times\frac{\frac{TP}{TP+FP}\times\frac{TP}{TP+FN}}{\frac{TP}{TP+FP}+\frac{TP}{TP+FN}} \quad (1)$$

Four additional measures were computed and analyzed in [14] but are not detailed here for space reasons.

Since the first step of a staff removal system is usually the detection of the staff lines, the overall performance highly depends on the accuracy of this preliminary staff detection. Indeed, if the staff line system is not detected (*i.e.* the image is rejected by the system), then the staff lines cannot be removed. Therefore, we compute the F-measures *with rejection*, *i.e.* for any rejected image, we consider that every staff line pixel is a False Negative and every non-staff line pixel is a True Negative (correctly classified as a non-staff line pixel). Furthermore, we also provide the number of rejected pages.

D. Performance Comparison

Figure 5 presents the F-measures obtained by the eight participant methods (see section III-A), and the baseline method (see section III-B). These results were computed from each of the 3 Test Subsets and 11 levels of degradation (see section II-C). For more details, please refer to [14].

The numbers on top of the bars correspond to the number of rejected images for a given method and a given level of degradation of each TestSubset. When no number is given, it means that the corresponding method does not reject any image on the corresponding subset. From these numbers, we can see that the NUASI-bin-lin and NUASI-bin-skel methods often reject some images (especially in the presence of 3D distortion and Mesh #2), but in a relatively low number (respectively 21 and 18 in total). On the other hand, the INESC-gray method rejects a higher number of images (80 in total) in the presence of a combination of 3D distortion and local noise. This is certainly due to the weighting function or binarization, as its binary version INESC-bin does not reject any image.

Concerning the 3D distortion (TestSubsets 1 and 3), most methods are less robust to perspective deformation (Mesh #1) than to the presence of small curves and folds (Mesh #2), except LRDE-gray, INESC-bin and INESC-gray which perform better in the presence of Mesh #1.

In the presence of local noise, the average F-measure (over the nine methods) on TestSubset2 drops of almost 4% when the level of noise increases from Low to High.

When combining local noise and 3D distortions (TestSubset3), the robustness of most methods drops drastically. Indeed, when mixing a high level of local noise and 3D distortions, the average F-measure (over the nine methods) drops of almost 6% compared to a high level of local noise only (for both meshes), and it decreases of respectively 3.88% and 3.22%, compared to 3D distortions with Meshes #1 and #2.

From Fig. 5, we can conclude that INESC-bin is the best on TestSubset2 containing local noise, while LRDE-bin is the winner on TestSubsets 1 and 3, containing respectively 3D distortions and a combination of 3D distortions and local noise.

IV. Conclusion

The second music scores competition on staff removal held in ICDAR2013 has raised a great interest from the research community, with 8 participant methods in the competition. For this competition, we generated a database of 6000 semi-synthetic images using the 1000 images from the CVC-MUSCIMA database and two models of degradation specifically designed to mimic the defects that can be seen in historical documents. This database contains 3 subsets both for training and testing, with separate and combined degradations, and in total 11 different levels of degradation. The performances of the 8 methods proposed by the 5 participants are analyzed and compared to a baseline method relying on the analysis of neighbouring connected components.

The methods submitted by all participants have obtained satisfying performance, even though the degradations in the proposed images are severe. But, most methods significantly decrease their performance when dealing with a higher level

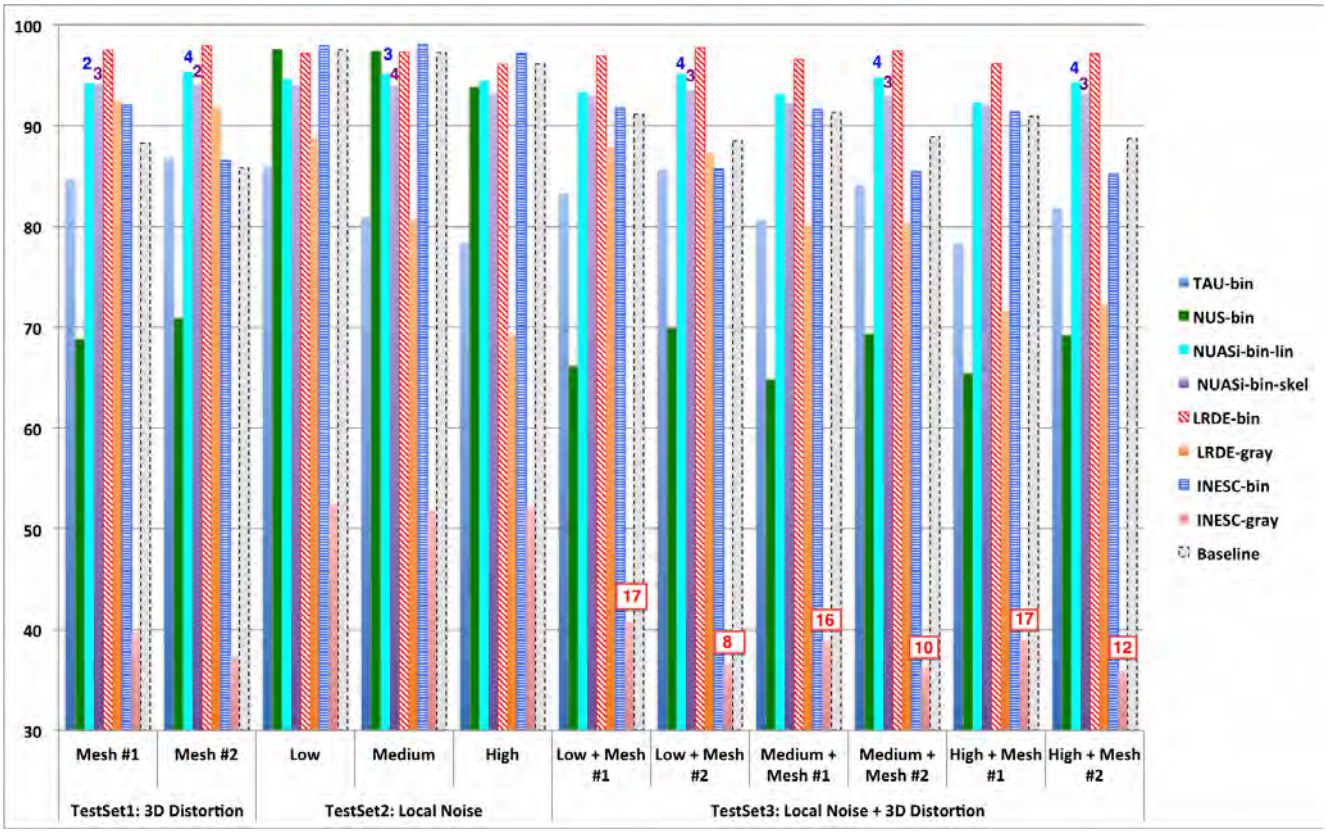


Fig. 5. F-measures of the eight participant methods (plus the baseline method) on the 3 Test Subsets and 11 levels of degradation.

of degradation, especially when combining 3D distorsions and local noise. The semi-synthetic ICDAR/GREC 2013 database is now available on the internet and labelled with different types and levels of degradation for both the training set and the test set; we hope the community will adopt it as a benchmark database for the research on handwritten music scores.

ACKNOWLEDGEMENTS

This research was partially funded by the French National Research Agency (ANR) via the DIGIDOC project, the spanish projects TIN2009-14633-C03-03 and TIN2012-37475-C02-02. We thank Anjan Dutta for providing the baseline results.

REFERENCES

[1] D. Blostein and H. S. Baird, *Structured Document Image Analysis*. Springer Verlag, 1992, ch. A Critical Survey of Music Image Analysis, pp. 405–434.

[2] A. Rebelo, I. Fujinaga, F. Paszkiewicz, A. Marcal, C. Guedes, and J. Cardoso, "Optical Music Recognition: State-of-the-Art and Open Issues," *International Journal of Multimedia Information Retrieval*, vol. 1, no. 3, pp. 173–190, 2012.

[3] C. Dalitz, M. Droettboom, B. Pranzas, and I. Fujinaga, "A Comparative Study of Staff Removal Algorithms," *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 30, no. 5, pp. 753–766, 2008.

[4] J. dos Santos Cardoso, A. Capela, A. Rebelo, C. Guedes, and J. Pinto da Costa, "Staff Detection with Stable Paths," *IEEE Trans. on PAMI*, vol. 31, no. 6, pp. 1134–1139, 2009.

[5] A. Fornés, A. Dutta, A. Gordo, and J. Lladós, "The ICDAR 2011 Music Scores Competition: Staff Removal and Writer Identification," in *International Conference on Document Analysis and Recognition (ICDAR)*, Beijing, China, September 2011, pp. 1511–1515.

[6] ——, "The 2012 Music Scores Competitions: Staff Removal and Writer Identification," in *Graphics Recognition. New Trends and Challenges. Lecture Notes in Computer Science*, Y.-B. Kwon and J.-M. Ogier, Eds. Springer, 2013, vol. 7423, pp. 173–186.

[7] ——, "CVC-MUSCIMA: a Ground Truth of Handwritten Music Score Images for Writer Identification and Staff Removal," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 15, no. 3, pp. 243–251, 2012.

[8] V. Kieu, N. Journet, M. Visani, R. Mullot, and J. Domenger, "Semi-synthetic Document Image Generation Using Texture Mapping on Scanned 3D Document Shapes," in *ICDAR 2013*, Washington, DC, USA, Accepted paper.

[9] V. Kieu, M. Visani, N. Journet, J. P. Domenger, and R. Mullot, "A Character Degradation Model for Grayscale Ancient Document Images," in *21st International Conference on Pattern Recognition (ICPR)*, Tsukuba Science City, Japan, Nov. 2012, pp. 685–688.

[10] I. Fujinaga, "Adaptive Optical Music Recognition," PhD Thesis, McGill University, 1996.

[11] B. Su, S. Lu, U. Pal, and C. L. Tan, "An Effective Staff Detection and Removal Technique for Musical Documents," in *10th IAPR International Workshop on Document Analysis Systems (DAS)*, Gold Coast, Queensland, Australia, March 2012, pp. 160–164.

[12] J. Cardoso and A. Rebelo, "Robust Staffline Thickness and Distance Estimation in Binary and Gray-Level Music Scores," in *20th International Conference on Pattern Recognition (ICPR)*, Istanbul, Turkey, August 2010, pp. 1856–1859.

[13] A. Dutta, U. Pal, A. Fornés, and J. Lladós, "An Efficient Staff Removal Approach from Printed Musical Documents," in *ICPR*, Istanbul, Turkey, August 2010, pp. 1965–1968.

[14] V. Kieu, A. Fornés, M. Visani, and N. Journet, "The ICDAR/GREC 2013 Music Scores Competition on Staff Removal," in *10th IAPR International Workshop on Graphics RECognition (GREC 2013)*, Bethlehem, PA, USA, Submitted paper.