The ICDAR/GREC 2013 Music Scores Competition: Staff Removal

Alicia Fornés[1], V.C. Kieu[2,3], Muriel Visani[2], Nicholas Journet[3], and Anjan Dutta[1]

[1]Computer Vision Center - Dept. of Computer Science,
Universitat Autònoma de Barcelona, Ed.O, 08193, Bellaterra, Spain
[2] Laboratoire Informatique, Image et Interaction - L3i
University of La Rochelle, La Rochelle, France
[3] Laboratoire Bordelais de Recherche en Informatique
LaBRI, University of Bordeaux I, Bordeaux, France
afornes@cvc.uab.es, muriel.visani@univ-lr.fr, {vkieu, journet}@labri.fr

Abstract. The first competition on music scores that was organized at ICDAR and GREC in 2011 awoke the interest of researchers, who participated in both staff removal and writer identification tasks. In this second edition, we focus on the staff removal task and simulate a real case scenario concerning old and degraded music scores. For this purpose, we have generated a new set of semi-synthetic images using two degradation models that we previously introduced: local noise and 3D distortions. In this extended paper we provide an extended description of the dataset, degradation models, evaluation metrics, the participant's methods and the obtained results that could not be presented at ICDAR and GREC proceedings due to page limitations.

1 Introduction

The recognition of music scores has been an active research field for decades [1, 2]. Many researchers in Optical Music Recognition have proposed staff removal algorithms in order to make easier the segmentation and enhance the accuracy of music symbol recognition [3, 4]. However, the staff removal task cannot be considered as a solved problem, especially when dealing with degraded handwritten music scores. This task is even defined as one of the "three challenges that should be addressed in future work on OMR as applied to manuscript scores" in the 2012 survey of Rebelo et al. [2].

At ICDAR /GREC 2011, we organized the first edition of the music scores competition [5]. For the staff removal task, we created several sets of distorted images in order to test the robustness of the staff removal algorithms. Each set corresponded to a different kind of distortion (e.g. Kanungo noise, rotation, curvature, staffline interruption, typeset emulation, staffline y-variation, staffline thickness ratio, staffline thickness variation and white speckles). The staff removal task woke up the interest of researchers, with eight participant methods.

Most staff removal methods showed good performance in front of severe distortions, although the detection of the staff lines still needed improvement.

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 one single kind of distortion. For this purpose, we combined some of the ICDAR 2011 distortions at different levels to create new sets of degraded images. We then asked the participants to run their algorithms on this new set of images. Unsurprisingly, the new results demonstrated that the performances of most methods were significantly decreased because of the combination of distortions.

By organizing a second edition of this competition, we aim at fostering the interest of researchers and focusing on the challenging problem of old document image analysis and recognition. For this second edition, we have generated realistic semi-synthetic images that emulate typical degradations appearing in old handwritten documents such as local noise and 3D distortions.

The rest of the paper is organized as follows. Firstly, we will describe the original dataset, the degradation models, and the generated training and test sets. Then, we will present the participants' methods. Finally, we will detail the the evaluation metrics, the results analysis, and conclude the paper.

2 Database

In this section we describe the original database, the degradation methods, and the semi-synthetic database for the competition.

2.1 Original CVC-MUSCIMA database

The original CVC-MUSCIMA [1] database [7] consists of 1,000 handwritten music score images, written by 50 different musicians. The 50 writers are adult musicians, in order to ensure that they have their own characteristic handwriting music style. Each writer has transcribed exactly the same 20 music pages, using the same pen and kind of music paper. The 20 selected music sheets contain music scores for solo instruments, choir or orchestra. For more information on the database (e.g. resolution, kind of music compositions, presence of text, etc.), the reader is referred to [7].

2.2 Degradation Models

3D Degradation Model Since the geometric distortions such as skews and curvatures are challenging for detecting staffs, we used them for the 2011 staff removal competition [5, 6]. However, these distortion models were only 2D models which are unable to reproduce the geometric distortions commonly encountered in real old documents such as dents, small folds, and torns... (see Fig. 1). Therefore, in this 2013 edition, we use the 3D degradation [8] that can generate more

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

realistic and more 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. It can 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 music score images on different 3D meshes. For more details, please refer to [8].

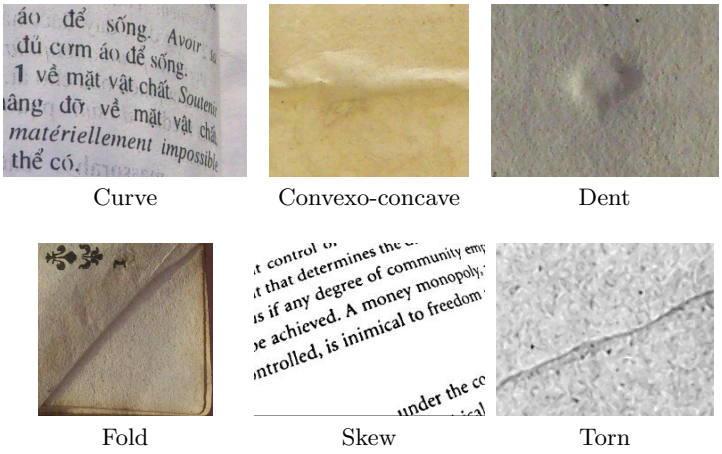

Fig. 1: Geometric distortions in real document images

Local Noise Model This model, introduced in [9], can mimic old documents' defects due to document aging and to the old printing/writing processes. Examples of these defects include ink splotches and white specks or streaks (see Fig. 2). Such defects might break the connectivity of strokes or add a connection between separate strokes. For staff line removal algorithms, local noise can lead to many types of challenging degradations. Indeed, it can lead to disconnections of the staff lines or to the addition of dark specks connected to a staff line. In the latter case, for instance, the dark specks might be confused with musical symbols.

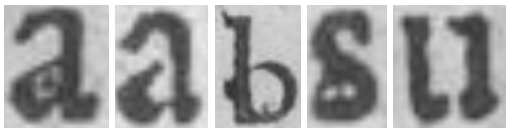

Fig. 2: Examples of local noise in real old documents

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 in the neighborhood of connected components' borders (obtained by binarizing the input grayscale image). Then, we define an arbitrary noise region at each seed-point (in our case, its shape is an ellipse). Finallly, the grey-level values of the pixels inside the noise regions are modified so as to obtain realistic looking bright and dark specks, and mimic defects due to the age of the document (ink fading, paper degradation...) and writing process (ink drops).

2.3 Degraded Database

For comparing the robustness of the staff removal algorithms proposed by the participants to this competition, we degrade the original CVC-MUSCIMA database using the two degradation models presented in Section 2.2. As a result, we obtain a semi-synthetic database that consists of 4000 images in the training set and 2000 images in the test set.

Training Set It consists in 4000 semi-synthetic images generated from 667 out of the 1000 original images. This set is split into the three following subsets:

- *TrainingSubset1* (see Fig. 3) contains 1000 images generated using the 3D distortion model and two different meshes. The first mesh consists in a perspective distortion due to the scanning of a thick and bound volume, while the second one contains many small curves, folds and concavities. We wrap the 667 original images on two meshes to produce 2x667=1334 semi-synthetic images. Then, 500 images per mesh are randomly selected so as to obtain a total of 1000 images.

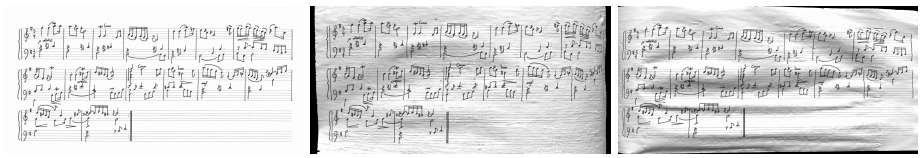

Fig. 3: TrainingSubset1 samples. From left to right: original image and two semi-synthetic images generated from two different meshes.

- *TrainingSubset2* (see Fig. 4) contains 1000 images generated with three levels of local noise (see Sub-section 2.2), as follows (the flattening factor of the elliptic noise region is fixed as g=0.6 for the three levels, whereas the noise region size a_0 increases after each level):
  - *Low level*: 333 images, 500 seed-points, a_0: 7;
  - *Medium level*: 334 images, 1000 seed-points, a_0: 8.5;
  - *High level*: 333 images, 1300 seed-points, a_0: 10.

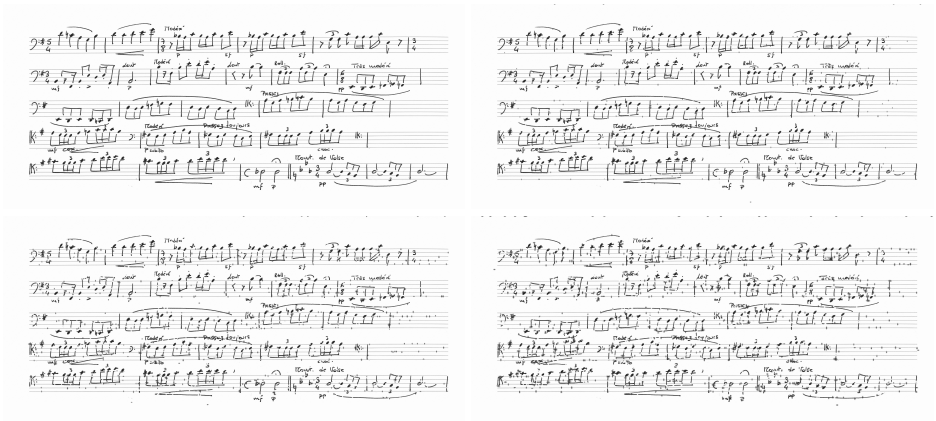

Fig. 4: TrainingSubset2 samples. From left to right and top to bottom: original image and semi-synthetic images generated from the original image using the low, middle and high levels of local noise.

– *TrainingSubset3* (see Fig. 5) contains 2000 images generated using both the 3D distortion and the local noise models. We obtain six different levels of degradation (the two meshes used for TrainingSubset1 $\times$ the three levels of distortion used for TrainingSubset2).

For each image in the training set, we provide its grey and binary versions. The associated ground-truth are the binary staff-less version (binary images without staff lines), as illustrated in Fig. 6.

Test Set It consists of 2000 semi-synthetic images generated from the 333 original images that differ from the ones for the training set.

– *TestSubset1* contains 500 images generated using the 3D distortion (see subsection 2.2). Two meshes - distinct from the ones used in the training set - are applied to the 333 original images and then only 500 images (250 for each mesh) are randomly selected among the 2x333=666 degraded images.
– *TestSubset2* contains 500 images generated using three levels of local noise, using the same values of the parameters as in TrainingSubset2, under the proportions $\frac{1}{3}$ / $\frac{1}{3}$ / $\frac{1}{3}$.
– *TestSubset3* contains 1000 images equally distributed between six different levels of degradation. These six levels of degradation come from the combination of the same two meshes as in TestSubset1 with the same three different levels of local noise as in TrainingSubset2.

For each image in the test set, we provide its grey and binary versions. The test set was provided to the participants 46 days after the training set (containing 4000 degraded images together with their ground-truth). The participants were

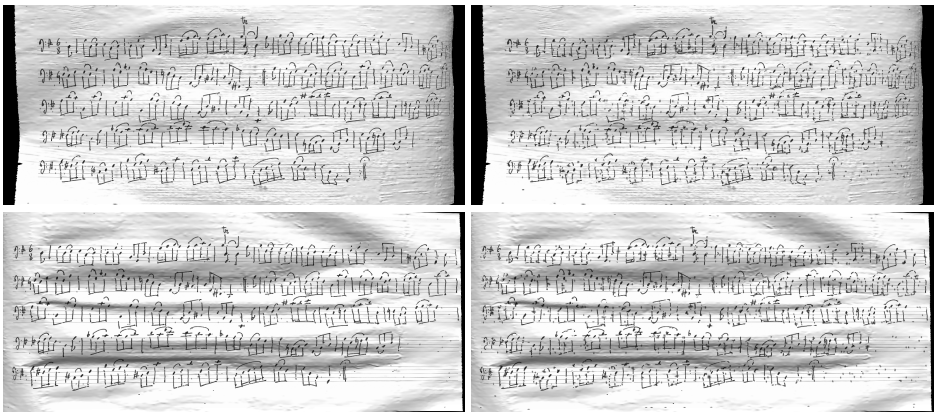

Fig. 5: TrainingSubset3 samples. First row, from left to right: images generated using mesh 1 and the low and high levels of local noise. Second row, from left to right: images generated using mesh 2 and the low and high levels of local noise.

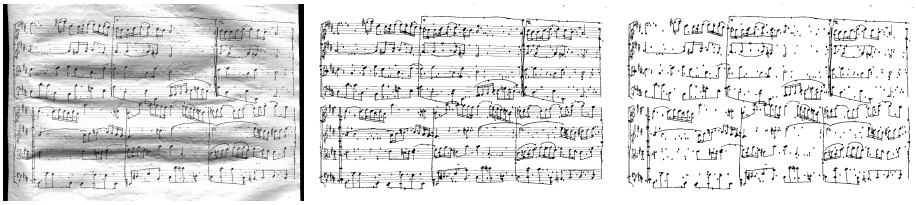

Fig. 6: From left to right: an image from TrainingSubset3, its binary version and its binary staff-less version (ground-truth).

asked to send to the organizers the outputs of their algorithms as binary staff-less images (such images containing only binarized music symbols but no staff lines) 23 days after the test set was provided to them.

3 Participants Information

In this section we will briefly describe the eight submitted methods of five 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.

3.1 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 globally deskewed, and the staff lines are located using a projection on the y-axis. Finally, the staff lines are removed using masks.

3.2 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 in the image. Finally, those pixels belonging to these staff lines are removed.

3.3 NUASi

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

- NUASi-bin-lin: This method is described in Section II of [3]. First, the staffs are detected, and 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 function *chordlength*(φ) (where φ is the angle of the chord at the intersection region) is used to filter staff-line pixels belonging to a crossing music symbol. 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. The staff segment results may contain false positive staff segments (*e.g.* in the case where a staff line is tangent with the curve of a music symbol or it overlaps with the music symbol at a staff segment). Then, to check for the false positives, non-staff segments which have the same splitting point as a staff segment are extrapolated by a parametric parabola. If the staff segment is tangent with the parabola, it is a non-staff segment. Finally, vertical black runs around the detected staff skeleton are removed when they are shorter than 2* *staffline_height*. The source code is available at *http://music-staves.sourceforge.net/(class MusicStaves_skeleton)*.

3.4 LRDE

Thierry Géraud, from the EPITA Research and Development Laboratory (LRDE), Paris, France, submitted two methods. For more details, the reader is referred to *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. The resulting image serves as a mask in which a two-level thresholding with hysteresis of the original image is applied. Finally, some spurious horizontal parts of the staff-lines are erased in a post-processing step.

3.5 INESC

Ana Rebelo and Jaime S. Cardoso (INESC Porto and Universidade do Porto) submitted the following two methods (more details are given in [4]) 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.
– INESC-gray: For grayscale images, the weight function is generalized by using a sigmoid function. The parameters of the sigmoid function are chosen to favor the luminance levels of the stafflines. A state-of-the-art binarization technique is used in order to assign the cost for each pixel in the graph (pixels binarized to white have a high cost; pixels binarized to black have a low cost). Once the image is binarized, the previous method is applied.

4 Results

In this section we compare the participant's output images with the ground-truth (binary staff-less images) of the test set using the measures presented in the next Section. The ground-truth associated to the test set was made public after the competition.

4.1 Measures Used for Performance Comparison

The staff removal problem is considered as a two-class classification problem at the pixel level. For each test subset and each level of noise, we compare the output images provided by the participants to their corresponding ground-truth. For this purpose, we compute the number of True Positive pixels (TP, pixels correctly classified as staff lines), True Negative pixels (TN, pixels correctly classified as non-staff lines) False Positive pixels (FP, pixels wrongly classified as staff lines) and False Negative pixels (FN, pixels wrongly classified as non-staff lines). Then, from these measures, we compute the accuracy (also called Classification Rate), precision (also called Positive Predictive Value), recall (also called True Positive Rate or sensitivity), F-measure and specificity (or True Negative Rate) as follows:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$precision = \frac{TP}{TP + FP} \tag{2}$$

$$recall = \frac{TP}{TP + FN} \tag{3}$$

$$F - Measure = 2 \times \frac{precision \times recall}{precision + recall} \tag{4}$$

$$specificity = \frac{TN}{TN + FP} \tag{5}$$

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 is undetected, it is unable to be removed. For example, a staff removal system may obtain very good results (when the staff is correctly detected) while rejecting many music scores images (if no staff line is detected in the image, it is discarded). Therefore, for each participant's method, we provide the number of rejected pages for each of the three test subsets, and for each level of degradation inside each test subset. Furthermore, we compute the evaluation measures (1-5) in two ways:

- Without rejection: the five average values of the five measures (1-5) are computed inside each test subset and for each level of degradation, taking into account only the images that the system pre-classified as music score. Thus, the rejected images are not taken into account for these measures.
- With rejection: the five average values of the five measures (1-5) are computed inside each test subset and for each level of degradation, taking into account every image in the test set, no matter if it was rejected by the system or not. Thus, for a rejected image, every staff line pixel is considered as a False Negative and every non-staff line pixel is considered as a False Positive.

4.2 Baseline

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

4.3 Performance comparison

Table 1 and Fig. 7 present staff removal results obtained by the eight participant methods (three out of the five participants presented two different algorithms). We have also tested the baseline method presented in [13]. Each of the nine columns indicates the name of the participant and the staff removal method category. All the metrics used for performance comparison (presented in section 4.1) are presented according to the type of degradation (3D distortion, local noise model, or a mixture of the two) and the degree of degradation. Indeed, for each sub-dataset five metrics are presented. The winner's method of each line is highlighted in bold.

Since the precision is higher in some methods but with a lower recall, we decided to select the winners according to the accuracy and the F-measure metrics shown in Fig. 7. From Table 1, we can see that the LRDE-bin method is the winner of the 3D distortion set, the INESC-bin is the winner of the Local noise set, and LRDE-bin is the winner of the combination set. Fig. 7 presents averages Accuracy and F-Measures of the nine tested methods on three different sub-datasets (3D distortion, local noise model, a mixture of the two), whereas

Table 1: Competition results for each test subset and each degradation level. We give the number # of rejected images, and the values of the measures computed with and without rejection. The measures (M.), showed in %, are: P=Precision, R=Recall, F-M=F-Measure, S=Specificity, A=Accuracy.

Deform.	Level	M.	TAU-bin	NUS-bin	NUASi-bin-lin	NUASi-bin-skel	LRDE bin	LRDE gray	INESC bin	INESC gray	Baseline
Set1: 3D dist.	Mesh 1 (M1)	P	75.51	98.75	99.05	98.58	98.89	87.26	**99.76**	32.50	98.62
		R	96.32	52.80	#2 $89.90_{(89.77)}$	#3 $90.26_{(90.03)}$	96.19	**98.41**	85.41	50.91	79.86
		F-M	84.65	68.81	$94.25_{(94.18)}$	$94.24_{(94.11)}$	**97.52**	92.50	92.03	39.67	88.26
		S	98.81	99.97	$99.96_{(99.96)}$	$99.95_{(99.95)}$	97.52	99.45	**99.99**	95.97	99.95
		A	98.721	98.25	$99.60_{(99.60)}$	$99.60_{(99.60)}$	**99.82**	99.42	99.46	94.32	99.22
	Mesh 2 (M2)	P	82.22	99.50	99.70	99.39	99.52	86.59	**99.90**	34.36	99.29
		R	91.90	55.05	#4 $92.07_{(91.38)}$	#2 $89.63_{(89.36)}$	96.39	**97.76**	76.33	40.85	75.47
		F-M	86.79	70.88	$95.73_{(95.36)}$	$94.26_{(94.11)}$	**97.93**	91.83	86.54	37.33	85.76
		S	99.26	99.99	$99.98_{(99.99)}$	$99.97_{(99.97)}$	99.98	99.44	**99.99**	97.12	99.98
		A	99.01	98.39	$99.71_{(99.68)}$	$99.61_{(99.60)}$	**99.86**	99.38	99.16	95.12	99.10
Set2: Local Noise	High (H)	P	65.71	95.37	**98.41**	97.28	95.54	53.22	97.63	38.81	95.65
		R	97.01	92.27	90.81	89.35	96.65	**98.58**	96.62	79.35	96.53
		F-M	78.35	93.79	94.46	93.15	96.09	69.12	**97.13**	52.13	96.09
		S	98.59	99.87	**99.95**	99.93	99.87	97.58	99.93	96.51	99.87
		A	98.55	99.67	99.71	99.64	99.79	97.61	**99.85**	96.05	99.78
	Medium (M)	P	69.30	97.82	**99.24**	98.38	97.50	68.10	98.95	39.61	97.26
		R	97.34	96.97	#3 $91.94_{(91.41)}$	#4 $90.56_{(89.80)}$	97.13	**98.77**	97.19	74.83	97.10
		F-M	80.96	97.39	$95.45_{(95.16)}$	$94.31_{(93.90)}$	97.32	80.62	**98.07**	51.81	97.18
		S	98.71	99.93	$99.97_{(99.97)}$	$99.95_{(99.95)}$	99.92	98.61	**99.96**	96.58	99.91
		A	98.67	99.85	$99.75_{(99.73)}$	$99.68_{(99.66)}$	99.84	98.62	**99.89**	95.96	99.83
	Low (L)	P	77.07	98.56	99.25	98.07	97.89	80.65	**99.42**	40.13	98.52
		R	96.88	96.58	90.48	90.17	96.47	**98.47**	96.52	75.48	96.45
		F-M	85.85	97.56	94.66	93.95	97.17	88.67	**97.95**	52.40	97.47
		S	99.12	99.95	**99.97**	99.94	99.93	99.28	99.98	96.59	99.95
		A	99.06	99.86	99.70	99.66	99.84	99.26	**99.88**	95.98	99.85
Set3: 3D dist. + Local Noise	H + M1	P	66.01	94.31	96.88	96.37	96.14	56.19	**97.63**	31.70	96.41
		R	96.35	50.00	88.03	87.93	96.13	**98.59**	85.79	#17 $55.21_{(50.48)}$	85.98
		F-M	78.34	65.35	92.25	91.96	**96.14**	71.58	91.33	$40.27_{(38.94)}$	90.90
		S	98.30	99.89	99.90	99.88	99.86	97.37	**99.92**	$95.93_{(96.27)}$	99.89
		A	98.24	98.25	99.51	99.49	**99.74**	97.41	99.46	$94.58_{(94.76)}$	99.43
	H + M2	P	73.40	97.50	**98.55**	98.07	97.61	57.18	98.35	33.11	97.62
		R	92.42	53.56	#4 $90.99_{(90.32)}$	#3 $89.15_{(88.68)}$	96.66	**98.00**	75.17	#12 $42.15_{(39.19)}$	81.26
		F-M	81.82	69.14	$94.62_{(94.25)}$	$93.40_{(93.14)}$	**97.13**	72.22	85.22	$37.09_{(35.90)}$	88.69
		S	98.86	**99.95**	$99.95_{(99.95)}$	$99.94_{(99.94)}$	99.92	97.51	99.95	$97.11_{(97.31)}$	99.93
		A	98.65	98.43	$99.66_{(99.64)}$	$99.59_{(99.57)}$	**99.81**	97.53	99.14	$95.31_{(95.41)}$	99.32
	M + M1	P	69.26	95.45	97.52	96.93	97.11	67.44	**98.51**	32.34	97.29
		R	96.44	49.07	89.15	87.98	95.98	**98.46**	85.63	#16 $53.52_{(48.76)}$	85.96
		F-M	80.62	64.81	93.15	92.24	**96.54**	80.05	91.62	$40.31_{(38.88)}$	91.27
		S	98.47	99.91	99.91	99.90	99.89	98.30	**99.95**	$96.01_{(96.36)}$	99.91
		A	98.406	98.168	99.549	99.491	**99.763**	98.312	99.461	$94.556_{(94.730)}$	99.43
	M + M2	P	77.50	98.39	99.02	98.53	98.42	68.09	**99.06**	33.76	98.35
		R	91.83	53.47	#4 $91.57_{(90.85)}$	#3 $88.43_{(87.94)}$	96.52	**97.92**	75.21	#10 $41.64_{(39.13)}$	81.08
		F-M	84.05	69.29	$95.15_{(94.76)}$	$93.20_{(92.93)}$	**97.46**	80.27	85.50	$37.29_{(36.25)}$	88.88
		S	99.06	99.96	$99.96_{(99.96)}$	$99.95_{(99.95)}$	99.94	98.38	**99.97**	$97.12_{(97.30)}$	99.95
		A	98.87	98.39	$99.68_{(99.66)}$	$99.56_{(99.54)}$	**99.83**	98.37	99.13	$95.24_{(95.32)}$	99.31
	L + M1	P	73.28	96.75	98.06	97.50	97.92	79.32	**99.14**	32.77	97.96
		R	96.38	50.22	88.96	88.74	95.92	**98.38**	85.48	#17 $53.83_{(48.83)}$	85.23
		F-M	83.26	66.12	93.29	92.92	**96.91**	87.83	91.80	$40.74_{(39.22)}$	91.15
		S	98.70	99.93	99.93	99.91	99.92	99.05	**99.97**	$95.93_{(96.30)}$	99.93
		A	98.62	98.17	99.55	99.52	**99.78**	99.03	99.46	$94.44_{(94.62)}$	99.41
	L + M2	P	80.17	99.00	99.39	98.94	99.02	78.81	**99.53**	34.31	98.84
		R	91.98	54.01	#4 $91.97_{(91.22)}$	#3 $89.14_{(88.63)}$	96.46	**97.85**	75.18	#8 $41.34_{(39.08)}$	80.14
		F-M	85.67	69.89	$95.54_{(95.13)}$	$93.78_{(93.50)}$	**97.72**	87.30	85.66	$37.50_{(36.54)}$	88.52
		S	99.17	**99.98**	$99.97_{(99.98)}$	$99.96_{(99.96)}$	99.96	99.04	**99.98**	$97.13_{(97.28)}$	99.96
		A	98.92	98.37	$99.70_{(99.67)}$	$99.59_{(99.57)}$	**99.84**	99.01	99.12	$95.18_{(95.25)}$	99.27
Total rejected images			#0	#0	#21	#18	#0	#0	#0	#80	#0

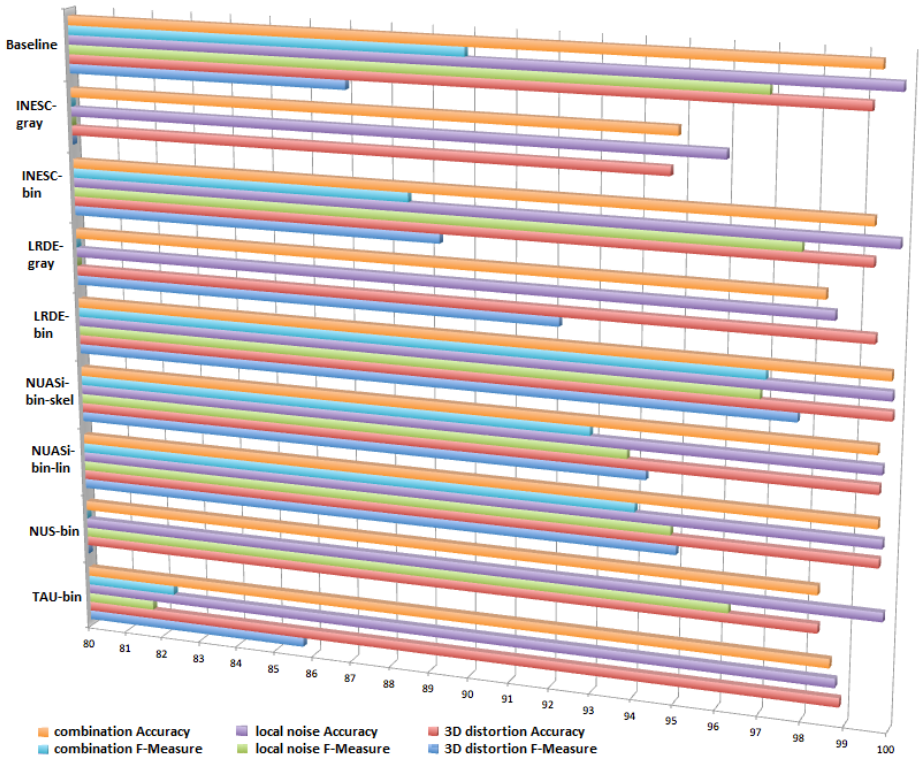

Fig. 7: Average music score competition results. The figure presents, for each participant method, the Average Accuracy and F-Measure of their method on the different sets.

Fig. 8 gives the values of the F-measures and the number of rejected images for each test subset and each level of degradation.

Much can be learned from the scores presented in Table 1. First, whatever the category (and intensity) of defect integrated in semi-synthetics images, the best results are mainly obtained with binary images (LRDE-bin, NUAS-bin-lin and INESC-bin). Only one method with grey-level image as input (LRDE-gray) gives 11 times the best results. It is also interesting that this method gives always the best recall score. The results showed that the baseline method [13] performances are almost identical to those obtained by each winning method.

Another interesting way to analyze these results is to compare the scores with the level of degradations of each sub dataset. Concerning 3D distortion, the staff lines images mapped on Mesh 1 (M1) were a bit more difficult to analyze for all the participants (except LRDE-gray and INESC-bin). On average the precision scores drop by 1 point. One method drops its precision by 7 points. We can conclude that these methods seems less robust to perspective deformation defects

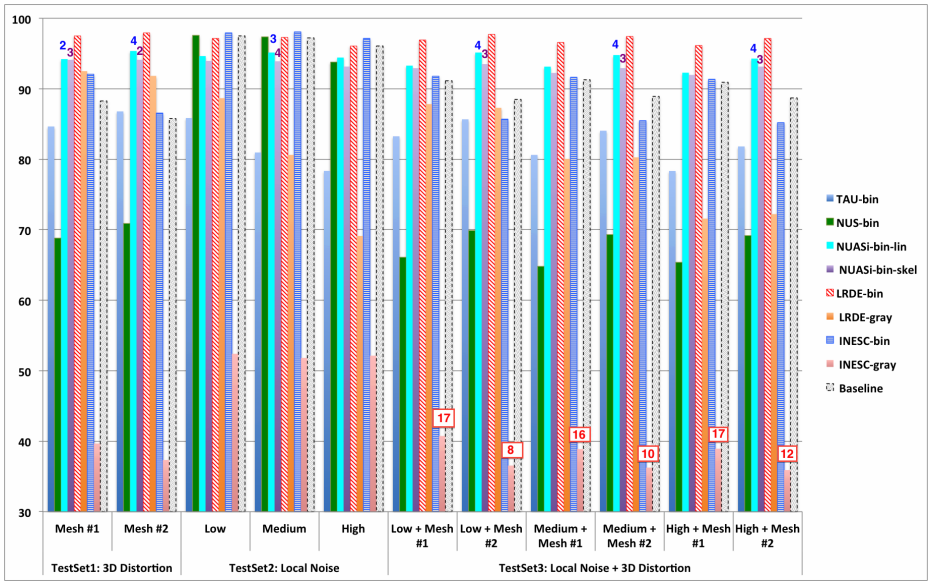

Fig. 8: F-measures of the eight participant methods (plus the baseline method) on the 3 Test Subsets and 11 levels of degradation.

(Mesh 1) than to the presence of small curves and folds (Mesh 2). The precision scores of each participant decrease when the local noise is getting higher. On average, the precision scores decrease from 13%. In the best case it dropped by less than 1%. In the worst case the precision is dropped by 33%. These results clearly show that all these methods are sensitive to the local noise deformation. The tests carried out with images generated by combining local noise and 3D distortions confirm that the results decrease when the level of degradation is important.

5 Conclusion

The second music scores competition on staff removal held in ICDAR2013 has raised a great interest from the research community, with eight participants' methods in the competition. For this competition, we generated a database of 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 three subsets both for training and testing: one subset containing only 3D distorsions at two different levels; one subset containing three levels of local noise and one subset with a combination of both sources of noises, with six different levels of degradation. We evaluated the performances of the eight methods proposed by the participants

and compared it towards a baseline method based on the analysis of neighbouring connected components.

The eight submitted methods have obtained very satisfying performances, even though the degradations in the proposed images were quite severe. The results of the participants have demonstrated that the performance of most methods significantly decreases when dealing with a higher level of degradation, especially in the presence of both sources of degradation (3D distortion model + local noise model). We hope that our semi-synthetic database, which is now available on the internet and labelled with different types and levels of degradation for both the training set and the test set, will become a benchmark for the research on handwritten music scores in the near future.

Acknowledgements

This research was partially funded by the French National Research Agency (ANR) via the DIGIDOC project, and the spanish projects TIN2011-24631 and TIN2012-37475-C02-02.

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 Transactions on Pattern Analysis and Machine Intelligence*, 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 Transactions on Pattern Analysis and Machine Intelligence*, 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)*, 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 *12th International Conference on Document Analysis and Recognition (ICDAR)*, 2013, pp. 489–493.
9. V. Kieu, M. Visani, N. Journet, J. P. Domenger, and R. Mullot, "A character degradation model for grayscale ancient document images," in *International Conference on Pattern Recognition (ICPR)*, Tsukuba Science City, Japan, Nov. 2012, pp. 685–688.

10. I. Fujinaga and B. Adviser-Pennycook, *Adaptive optical music recognition.* McGill University, 1997.
11. B. Su, S. Lu, U. Pal, and C. L. Tan, "An effective staff detection and removal technique for musical documents," in *IAPR International Workshop on Document Analysis Systems (DAS)*, 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)*, 2010, pp. 1856–1859.
13. A. Dutta, U. Pal, A. Fornés, and J. Lladós, "An efficient staff removal approach from printed musical documents," *International Conference on Pattern Recognition (ICPR)*, pp. 1965–1968, 2010.