Pattern Analysis and Applications manuscript No.
(will be inserted by the editor)

Avoiding staff removal stage in Optical Music Recognition - Application to scores written in White Mensural notation

Authors

Received: date / Accepted: date

Abstract Staff detection and removal is one of the most important issues in Optical Music Recognition tasks since common approaches for symbol detection and classification are based on this process. Due to its complexity, staff detection and removal is often inaccurate, leading to a great number of errors in posterior stages. For this reason, a new approach that avoids this stage is proposed in this paper, which is expected to overcome these drawbacks. Our approach is put into practice in a case of study focused on scores written in white mensural notation. Symbol detection is performed by using the vertical projection of the staves. The cross-correlation operator for template matching is used at the classification stage. The goodness of our proposal is shown in an experiment in which our proposal attains an extraction rate of 96 % and a classification rate of 92 %, on average. The results found have reinforced the idea of pursuing a new research line in OMR systems without the need of the removal of staff lines.

Keywords Optical Music Recognition · Staff Detection and Removal · Ancient Music · White Mensural Notation

1 Introduction

Since the emergence of computers, much effort has been devoted to digitizing music scores. This process facilitates music preservation as well as its storage, reproduction and distribution. Many tools have been developed for this purpose since the 1970s. One way of digitizing scores is to use electronic instruments (e.g. a MIDI piano) connected to the computer so that the musical information is directly transfered. However, this process is not free of errors and inaccuracies could cause differences between the generated score and the original one. An additional bothersome feature of this method is that it requires the participation of experts who know how to perform the musical piece. On the other hand, software for creating and editing digital scores, in which musical symbols are placed in a staff based on 'drag and drop' actions, are also available. Nevertheless, the transcription of scores with this kind of tools is a very time consuming task. This is why systems for automatic transcription of music scores became an important need.

Optical Music Recognition [1] (OMR) is the task of automatically extracting the musical information from an image of a score in order to export it to some digital format. A good review of OMR can be found in the work of Rebelo et al. [23], covering the state-of-the-art and the remaining challenges.

In this work, we are interested in the process of recognition of musical symbols from ancient scores. Ancient music is a main source of historical heritage. This kind of music is scattered across libraries, cathedrals and museums, what makes it difficult to access and study them. In order to use these documents without compromising their integrity, they can be digitized. However, conventional OMR systems are not effective transcribing ancient music scores [18]. The quality of the sheet, the inkblots or the irregular leveling of the pages constitute some features to overcome. Moreover, it is extremely complex to build systems for any type of document because several notations can be found: mensural (white and black), tablature, neumes, etc. In the literature, some studies that have worked with some kinds of ancient scores can be found, such as those reported in [19] or [8].

Institutions

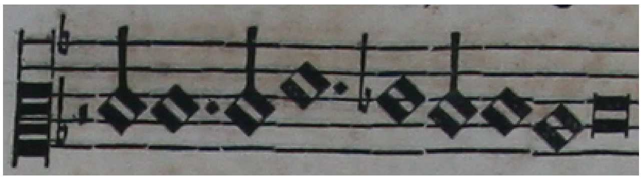


Fig. 1 Piece of staff in white mensural notation from the ACM. Each musical symbol is printed separately with its part of the staff.

The system described here focuses on analyzing ancient scores in white mensural notation. Specifically, our dataset consists of scores from the *Archivo de la Catedral de Malaga* (ACM). The ACM was created in the XV-th century and its library contains music scores from the X-th to the XX-th centuries. The scores of our dataset have a special feature: unlike other ancient printed scores in which the printing house put the symbols over an empty staff, these symbols were printed jointly with a piece of staff over an empty sheet (see Fig. 1). It means that a in each piece of the score, a single symbol is found. Furthermore, a noticeable distance between each musical symbol always exists. These features allow us to address the OMR process avoiding the common staff detection and removal stage.

Much research has been conducted in OMR concerning staff detection and removal [27, 25, 7]. This stage is one of the most critical aspects for both the detection and the classification of the musical symbols since they are based on symbol isolation. This stage is hardly sufficiently accurate and it often produces noisy results. Although more aggressive methods that minimize noise can be used, they produce partial or total loss of some musical symbols. The trade-off between these two aspects, in addition to the accuracy of the techniques, has hitherto led to the inevitable production of extraction and classification errors [23]. Furthermore, this stage is usually very expensive in terms of time. For this reason, other authors decided to face OMR without the staff removal stage. In the work developed in [16], the whole score (including the staff) is thinned by a skeleton algorithm. The symbols are then detected seeking junctions and termination points. Pugin [22] also proposed a recognition scheme in which the score maintains the staff lines. His approach consisted in learning Hidden Markov Models based on low-level.

Although these approaches are less common in the literature, we consider that this kind of procedure is an interesting option in different types of musical scores. Most of the current OMR systems are developed to handle contemporary notation but same algorithms are performed later to early music, which is characterized by different types of scores. In this work we propose an scheme that skips the staff removal stage. This approach is expected to helpt to reduce extraction and classification errors. Our aim is to show that this way of building OMR systems can be very effective for some music scores.

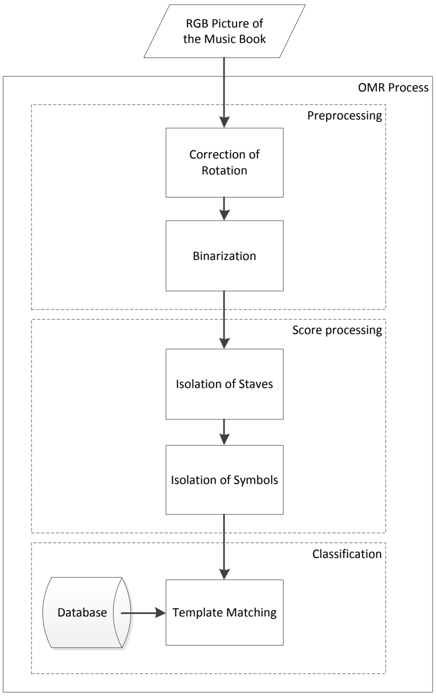


Fig. 2 General scheme of the recognition process.

The type of scores selected from the ACM give the possibility of detecting the musical symbols in a simple way. Since each symbol is on a different piece, there cannot be overlap. Therefore, in each piece of the score there can be only one symbol. The extraction of the musical symbols only requires the detection of the portions of the staff in which each symbol begins and ends. Moreover, keeping the staff lines forces us to select appropriate techniques to classify the musical pieces of symbols. In this paper, a method based on template matching is proposed, since all the symbols to be detected come from a fixed font type due to the engraving mechanism. This approach has been successfully used for OMR tasks in some previous works [30, 4].

The remaining paper is structured in the same way as the recognition process (see Fig. 2): Section 2 details the preprocessing stage, Section 3 describes the score processing task, in which each staff of the score is isolated and each symbol is detected, Section 4 presents the classification step. Results are shown in Section 5 and some conclusions are drawn in Section 6. The steps to be performed after the recognition of symbols will not be addressed. An example of those processes for scores written in white mensural notation can be found in [29].

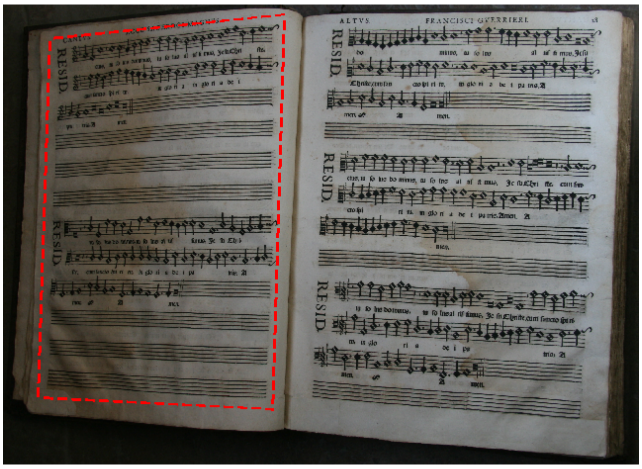


Fig. 4 Polygon over the ROI. The polygon identifies the boundaries of the page and provides the key points to correct the rotation.

2 Preprocessing Stage

In order to ensure the integrity of the documents, the images provided as input to the system correspond to pictures on polyphony books of the inventory of the ACM (Fig. 3), which consists of two pages each. A preprocessing of the image is a key step to perform the recognition task.

Often, the book appears rotated with respect to the image axes. Furthermore, the position of the book in the picture makes the perspective of the pages inconvenient. It is especially important to correct both the rotation and the perspective so that the musical symbols can be detected and recognized correctly. Also, the background of the pages and ink are acquired with different color levels depending on their location due to the sheet conditions (irregular leveling, uneven lighting, paper degradation, etc.). Therefore, a binarization process that allows distinguishing accurately between the background and ink seems crucial for the performance of the system as well as for reducing the complexity of the recognition. These two steps are considered in the next subsections.

2.1 Correction of Rotation

The process of transcription begins with the detection of the Region of Interest (ROI), which follows the same process as explained in [2]. The polygon that marks the boundaries of each page is found (Fig. 4). In addition to the separation of the pages, the vertexes of this polygon provide the key points to perform the correction of rotation.

The objective of this step is to correct the rotation of the page. A perfect alignment with the image axes constitutes the starting point for the following stages since they are based on the horizontal and vertical histograms to detect the different parts of interest. In the case of these images, it is not sufficient to perform a simple rotation because the pages (their projection in the image) do not have the shape of a rectangle, but a trapezoid. Thus, the rotation is corrected by recovering the perspective distortion of the image with respect to the book pages.

In order to perform this rotation we take the sides of the ROI polygon and split each pair into an equal number of segments to create a grid. Each pixel belonging to this grid is interpolated onto a rectangle. This process, when applied over a page of the input image, produces a result like the one shown in Fig. 5(a). It can be observed that both the alignment with the image axes and the perspective are now adjusted successfully.

2.2 Binarization

The next step of the preprocessing stage is to binarize the image. We should be able to distinguish between meaningful pixels (music symbols, staves) and others (background, mold, noise). However, the binarization cannot be applied directly to the image with a typical adaptive method because of the presence of irregularities in the sheet. Hence, the binarization requires a more comprehensive process. The actions needed to better perform the binarization of these sheets are:

- RGB to grayscale conversion: The input images are in RGB color space. Since the relevant information of each pixel for our task relies only on its position and its intensity, the image is converted to grayscale by using a weighted average [10].
- Contrast enhancement: In order to enhance the image, the Contrast-Limited Adaptive Histogram Equalization (CLAHE) algorithm [20] is applied.
- Illumination compensation: Since the illumination can vary largely among the set of images, the isolation of the reflectance –which keeps the meaningful information– is required. To this end, an aggressive symmetric Gaussian low-pass filter is used, so that an estimation of the illumination at each pixel can be obtained to correct the image. Preliminary experiments showed that a filter with size 80 and standard deviation 50 provided good results in the considered images. Nevertheless, results were not significantly different when using other similar parameters of the same order of magnitude.
- Adaptive thresholding: An adaptive method is now needed to find the threshold that clusters the background pixels and the pixels with ink. At this stage,

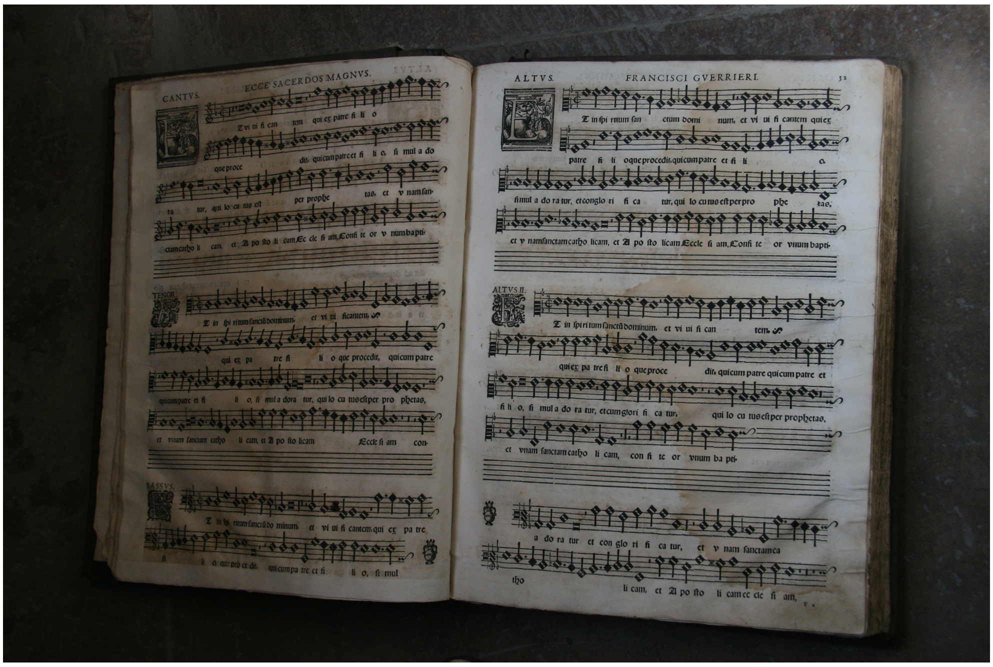
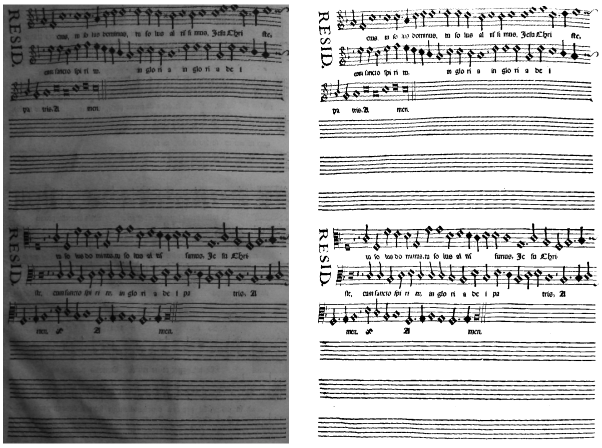


Fig. 3 Input image from the polyphony book 6 of the inventory of 1859 of the ACM (Francisco Guerrero, 1582).



(a) Perspective-corrected image

(b) Perspective-corrected image binarized by using Otsu's method

Fig. 5 Binarization of the perspective-corrected image.

the Otsu's method [17] –which is reported as one of the fastest and most successful algorithms for this purpose [31]– is finally used to binarize the image.

An example of the result of the binarization process can be found in Fig. 5(b).

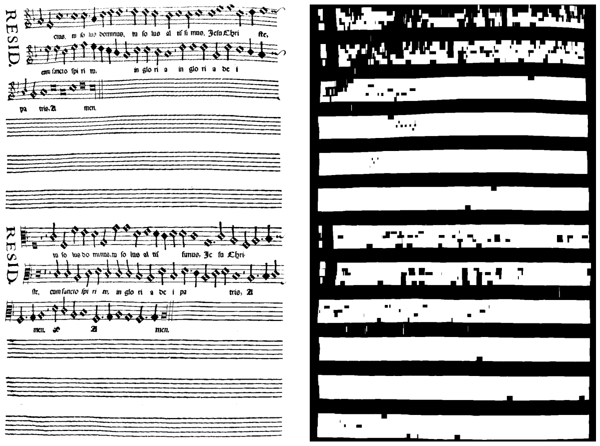
3 Score Processing

After the preprocessing stage, a binary image with perspective and rotation corrected is obtained. The next objective is to detect the musical symbols contained. As the scores are organized by staves, treating each staff separately is convenient. When the staves are isolated, the procedures for symbol detection can be performed more easily. In the next subsections, these two stages are described.

3.1 Isolation of Staves

Staff detection consists in seeking the positions of five equally-spaced parallel lines. The detection of the areas that contains these lines indicates the location of the staves. A common procedure is to compute the row histogram (or *y-projection*) of the image [28]. Staff features such as distance between staff lines, thickness of the staff lines and distance between staves are then computed from the histogram in order to isolate each staff. Alas, the presence in the scores of the ACM of other content such as lyrics or frontispieces among the staves complicates the process. Our approach handles this problem by creating a mask that keeps only the regions with horizontal lines. Unlike other works, we do not apply this mask to remove meaningless parts of the score but to directly isolate the staff parts on this mask.

First, an erosion over the binarized page is performed with a $1 - by - 20$ rectangular structuring element, which leads to the detection of parts with staves. A dilatation with a $20 - by - 1$ rectangular structuring element is then applied in order to span the entire space of the staff with the areas identified in the previous step. This way, a mask that indicates when a pixel is part of a staff region is estimated (Fig. 6).



(a) Binary image of the page (b) Mask over the staff regions

Fig. 6 Creation of a mask to detect staff regions.

It should be noted that by this mask, the extraction of staff features is not needed: staff splitting can now be performed with a row histogram analysis directly over the mask. Only a threshold is required in order to distinguish between rows with staff regions and rows with some remaining noise. Theoretically, each column of the histogram with a value higher than 1 should be considered part of a staff. Nevertheless, taking into account that previous steps are not error-free and staff parts get higher row-projection values, we decided to set a threshold which was a good margin with respect to the removal of noise and the detection of staff parts. Preliminary experiments established the threshold as 100 for the pages used in our experiments (1600 × 1000). This value achieved the best trade-off between noise removal and detection. Afterwards, the intersection of the threshold line with the slopes of the histogram indicates where each staff is located in the original image (Fig. 7).

3.2 Isolation of Symbols

After each staff has been isolated, the next goal is to detect the musical symbols contained. The common procedure at this point in typical OMR frameworks is the staff detection and removal. As aforementioned, we aim at exploring the possibilities of avoiding this step. The need of the removal of every part of the staff leads to delete some parts of the musical symbols, which produces unavoidable errors in posterior stages. Systems focused on contemporary scores need this process for the detection and classification of symbols. However, other scores –like the ones in our case– allow addressing the problem in a less aggressive manner and, eventually, less likely to delete important parts of the sheet. Thus, a novel approach for symbol detection and classification is presented.

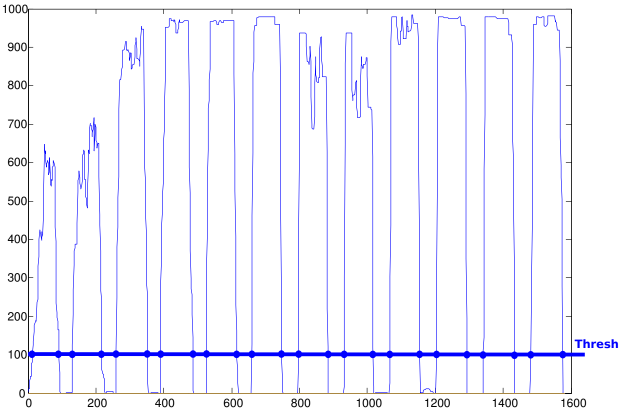


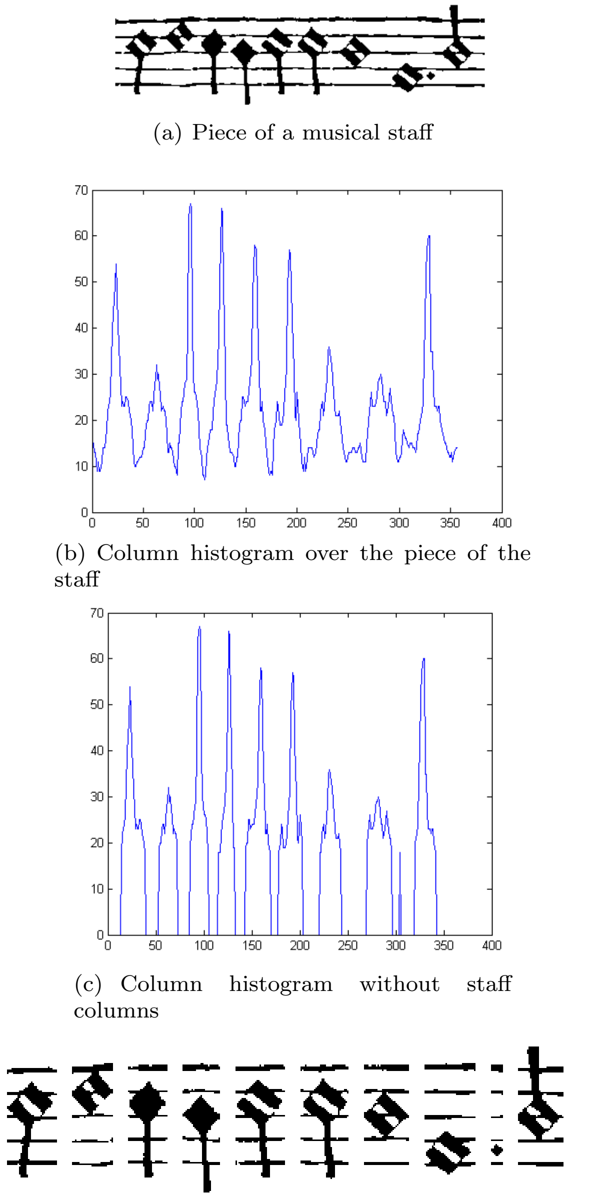
Fig. 7 Isolation of the staves. The intersection of the threshold line with the row histogram over the staff mask indicates the boundaries of each staff.

Instead of staff removal and detection, we directly extract the column histogram of each staff obtained in the previous section. This histogram contains enough information to detect the musical symbols. Over this histogram a k-means clustering [11], with $k = 3$, is applied to distinguish among the three column types considered: columns only with staff lines, columns with the head of a musical symbol, and columns with the head of a musical symbol and its stem. Manhattan distance [5] is used in the clustering method instead of the Euclidean because it has proven to be more accurate for our system. After this process, the cluster with the lowest centroid –that corresponds to the areas without musical symbols– is removed. The histogram found is then used to partition the staff. This process is illustrated in Fig. 8.

3.2.1 Special Staff Types

The process explained so far performs well for common staves. However, there are two types of staff in the ACM scores that require some specific attention: staves with frontispiece (Fig. 9(a)) and half-filled staves (Fig. 10(a)). The special features of these staves distort the results of the clustering process and can lead to a poor segmentation. A slight preprocessing stage for these staves is required.

In the first case, in order to prevent parts of the frontispiece being treated as musical symbols, the beginning of the staff should be detected. The column histogram



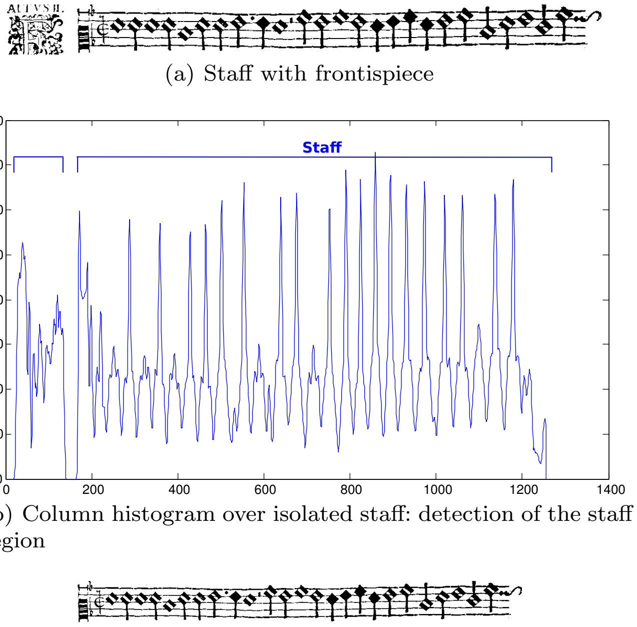
(a) Piece of a musical staff

(b) Column histogram over the piece of the staff

(c) Column histogram without staff columns

(d) Example of the extraction of musical symbols by histogram analysis

Fig. 8 Extraction of musical symbols from a piece of staff.

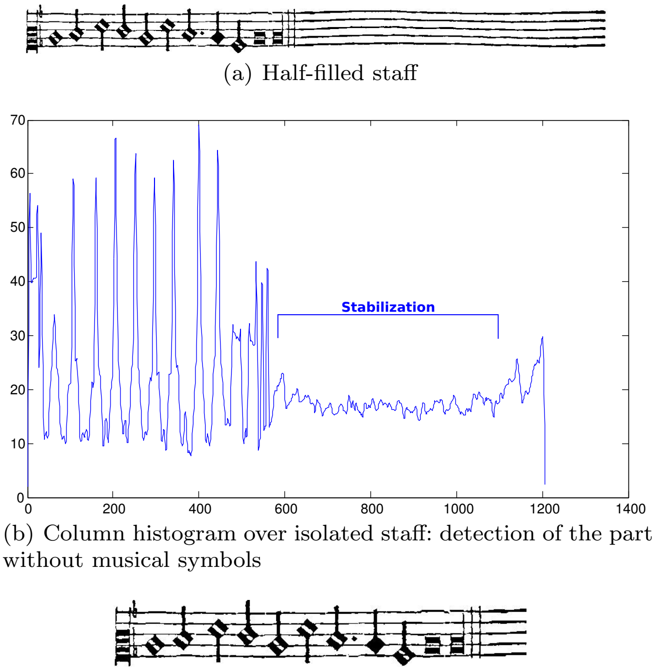


(a) Staff with frontispiece

(b) Column histogram over isolated staff: detection of the staff region

(c) Staff without frontispiece

Fig. 9 Preprocessing of a staff with frontispiece.



(a) Half-filled staff

(b) Column histogram over isolated staff: detection of the part without musical symbols

(c) Staff without the empty part

Fig. 10 Preprocessing of a half-filled staff.

is used to detect the connected parts and keep only the widest one, which is expected to correspond to the staff (see Fig. 9).

In the case of half-filled staves, a correct clustering of the columns without symbols is difficult to perform because the number of such columns represent a very large percentage with respect to the total number of columns to analyze. The solution to this problem is to trim the image so that the process is applied only to the parts that actually contain musical symbols. The detection of those parts is performed by means of a column histogram analysis. Starting from the left-hand side, it is checked if the histogram stabilizes within a meaningful period. If this happens, it can be assumed that the rest of the staff is empty so we trim the image at that point (see Fig. 10).

These two processes are applied to all the staves before the clustering process since they perform well regardless of the type of staff. It should be noted that only one vertical histogram is required to compute all the processes.

4 Classification

The output of the previous section is a set of ordered images containing a single musical symbol. The classification stage aims at labeling each of these images with the symbol contained in it. Typical OMR systems rely on feature extraction to classify the symbols. These features are then used to construct a set of samples to perform pattern recognition methods. Image feature extraction for recognition can be based on several techniques: Fourier descriptors [33], Angular-Radial Transform (ART) moments [12], chain codes such as Freeman's (FCCE) [9] or Vertex Chain Code (VCC) [3], etc. Unfortunately, these methods cannot be applied to these images as the presence of staff lines would represent an ineluctable obstacle. A classification method whose performance does not get severely damaged by the presence of the staff lines is required. This is the reason that led us to use the cross-correlation.

Cross-correlation [6] is a common method for template matching [24,32]. Let $f(x, y)$ be an image and $w(x, y)$ be a template, the cross-correlation can be computed with the following equation:

$$\gamma(u,v) = \frac{\sum_{x,y} [f(x,y) - \overline{w}_{u,v}][t(x-u,y-v) - \overline{w}]}{\sqrt[2]{\sum_{x,y} f(x,y) - \overline{f}_{u,v}]^2 [w(x-u,y-v) - \overline{w}]^2}} \tag{1}$$

where $\overline{f}_{u,v}$ is the mean of $f(x, y)$ in the region under the template and $\overline{w}$ is the mean of the template. Equation (1) is commonly referred to as *normalized cross-correlation* [26]. The result of the normalized cross-correlation gives a value between -1 and $+1$ related to the presence of the template at each point of the image. In this work, a fast version of the normalized cross-correlation [15] is used.

It should be noted that the cross-correlation matrix can give high values despite being different symbols as long as some piece of the image looks like the template. Fortunately, it is known that if there is a very high value in the center of the matrix, the probability of being the same symbol is very high. This is because all symbol images in our dataset contain the symbols centered horizontally. Thus, we establish that the correlation values of interest are those that are well centered horizontally. We assume that if the cross-correlation attains its maximum value close to the vertical edges, it should be considered a misclassification. Hence, the classification process is governed by a range $R = (x_s, x_e)$, normalized with respect to the width of the image ($x_s, x_e \in [0, 1]$), that indicates which cells of the cross-correlation matrix must be taken into account for the classification.

Let s represent the $N \times M$ image of a symbol, W stands for the dataset of labeled symbols, $L(w)$ represents the label of a template w; let $M(m)$ denote the maximum value of a matrix m, let $[m]_{a:b,c:d}$ represent the sub-matrix of m formed by rows $a, \ldots, b$ and columns $c, \ldots, d$, and let $R = (r_1, r_2)$ denote a specific range, with $r_1, r_2 \in [0, 1]$; the label τ of s (τ_s) is determined by the following equation:

$$\tau_s = L(\arg\max_{w \in W} M([\gamma(s,w)]_{[Nr_1:Nr_2,1:M]})) \tag{2}$$

In Eq. (2), the normalized cross-correlation between the extracted symbol and each labeled template in the database is applied. The template that achieves the best cross-correlation value within the width range R is used to label the symbol. It should be clear that, with this method, we can determine both the type and the pitch of the symbol as long as the labels in the database keep this information.

5 Experiments

In this section, some experiments are carried out to assess the accuracy of the proposed strategies. Our data set is composed of 12 pictures, with two pages each one. The average number of staves in each page is 12. Over the entire data set, 5768 symbols are to be extracted and classified. The parameters involved in the process are: the total number of musical symbols in the scores (T), the number of extracted symbols (E) and the number of correctly classified symbols (C). It should be noted that E can be divided into the number of musical symbols extracted (S_e) and the number of noise images extracted (N_e). All the symbols that either contain no musical information (e.g. parts of the frontispiece) or are partially (wrongly) extracted are considered as noise. Similarly, C can be divided into the number of correctly classified musical symbols (S_c) and the number of noisy symbols detected (N_c) –noise images classified as noise–.

Since the extraction and the classification are two different processes that can be evaluated separately, an evaluation for each process is performed. A global evaluation of the system, involving both the extraction and the classification, is also included.

5.1 Evaluation of the Extraction Process

A good performance of the symbol extraction stage is the first requirement to perform a good transcription. The extraction process is related to the number of musical symbols correctly extracted as well as to the number of symbols lost or partially (wrongly) extracted. In order to assess this process, we use the extraction rate. This parameter can be calculated as the number of musical symbols that have been found during the segmentation process divided by the total number of musical symbols in the score:

$$R_{\text{ext}} = \frac{S_e}{T} \tag{3}$$

Moreover, it is also important to quantify the noise introduced during the segmentation. The amount of noise can be evaluated by using the noise rate, based on the number of noise images extracted (N_e) and the total number of symbols extracted from the scores (E):

$$R_{\text{noise}} = \frac{N_e}{E} = \frac{N_e}{S_e + N_e} \tag{4}$$

Table 1 shows the extraction performance over our set of images. These results show that our extraction stage is able to achieve a rate over a 95 %, on average. All cases exceed a 93 %, even some of them are over 97 %. Moreover, the noise rate is low in almost all the cases, which means that our strategy accurately distinguishes between musical symbols and other objects of the scores. These values show the good performance of our symbol detection strategy.

Further analysis of these results revealed that the musical symbol *dot* is the most commonly missed symbol. The small width of the symbol makes it difficult to be detected. Changing the detection parameters so that this symbol gets detected more accurately led to a larger noise rate. We consider that it is preferable to accept some *dot* misses rather than generate a more noisy output which may deteriorate the whole transcription process.

Fold	T	S_e	N_e	R_{ext} (%)	R_{noise} (%)
1	390	371	3	95.13	0.80
2	377	361	7	95.76	1.90
3	623	598	5	95.99	0.83
4	432	421	10	97.45	2.32
5	410	399	2	97.32	0.50
6	427	414	8	96.96	1.90
7	514	498	7	96.89	1.39
8	436	425	6	97.48	1.39
9	441	433	3	98.19	0.69
10	444	432	5	97.30	1.14
11	633	598	9	94.47	1.48
12	641	601	7	93.76	1.15
Whole	5768	5551	72	96.24	1.28

Table 1 Performance results of the extraction process over the data set. The table contains information about the number of musical symbols in each fold (T), the number of musical symbols extracted (S_e) and the number of noise images extracted (N_e), which are used to calculate the extraction rate (R_{ext}) and the noise rate (R_{noise}).

5.2 Evaluation of the Classification Process

The evaluation of the classification process aims at measuring the goodness of the method used to determine the type of the symbols found. As indicated in Section 4, the cross-correlation operator for template matching was chosen. In our system, we evaluate the accuracy of the classification strategy regardless of the type of symbols detected or the type of error made, so, in order to evaluate the performance, we use the common $0 - 1$ loss function. This function is able to measure the rate of misclassified symbols if a uniform weight for each symbol is established. Thus, the classification rate can be defined as the number of correctly classified symbols divided by the number of symbols extracted:

$$R_{\text{classification}} = \frac{C}{E} = \frac{S_c + N_c}{S_e + R_e} \tag{5}$$

The classification experiment is conducted by using a k-fold cross validation scheme. Each fold is composed of one of the images of the data set while the labeled symbols of the rest of the folds are used as database for the cross-correlation operator. The results for each fold are shown in Table 2. A set of possible values for the range $R = (r_1, r_2)$ (Eq. (2)) are confronted experimentally.

The results show that the classification rate obtained with the cross-correlation is larger than 90 % in all the cases considered. Also, it has been shown that the best range to use for the cross-correlation is between 30 % and 70 % of the total width of the image, which yields a classification rate of 91.64 %, on average. However, it should be emphasized that the results among the different alternatives are not particularly remarkable, which is indicative of the robustness of the cross-correlation operator with respect to this parameter.

5.3 Global Evaluation

In the previous subsections, the extraction strategy and the classification strategy were evaluated. However, the OMR system has to be globally evaluated by involving both the extraction and the classification stages. In order to assess its performance, we use the well-known Word Error Rate (WER) [13].

Classification rate ($R_{\text{classification}}$)

		Range $R = (r_1, r_2)$				
Fold	E	(0,1)	(0.1,0.9)	(0.2,0.8)	(0.3,0.7)	(0.4,0.6)
1	374	92.25	93.04	92.78	93.85	93.58
2	368	88.86	89.40	91.30	91.30	91.84
3	603	88.22	88.39	88.55	88.72	88.05
4	431	91.18	91.87	92.34	93.03	93.27
5	401	91.52	91.52	93.01	93.26	93.76
6	422	91.23	90.75	92.18	92.41	92.65
7	505	89.30	89.50	91.28	91.48	90.89
8	431	89.32	89.79	91.41	91.41	91.18
9	436	92.66	92.88	92.88	93.11	93.80
10	437	88.55	88.55	91.99	92.67	93.13
11	607	89.12	88.96	89.45	89.45	89.29
12	608	90.29	90.78	90.78	91.44	91.11
Whole	5623	90.09	90.39	91.30	91.64	91.62

Table 2 Classification rate over the data set with a 12-fold cross validation scheme. Different ranges R for the cross-correlation are presented.

The WER is based on the edit distance [14] and measures the difference between two sequences (in our case, two sequences of musical symbols). As the focus of OMR systems is to assist the human task, this metric can provide an estimation of the human effort needed to correct the output of the system. It involves the three common edit operations, which in this case are defined as follows:

- Insertions: The difference between the number of musical symbols in the score and the number of extracted symbols ($T - S_e$)
- Substitutions: The difference between the number of extracted symbols and the number of symbols correctly classified ($S_e - S_c$)
- Deletions: The difference between the number of noise image extracted and the number of noise correctly classified ($N_e - N_c$)

Therefore, the WER can be calculated by summing up these three values and dividing it by the total number of musical symbols:

$$\begin{aligned} WER &= \frac{(T - S_e) + (S_e - S_c) + (N_e - N_c)}{T} \\ &= \frac{T + N_e - C}{T} \end{aligned} \tag{6}$$

The final results of applying our OMR process over the data set with the best classification parameter selected ($R = (0.3, 0.7)$) are shown in Table 3. Note that since we are reporting the accuracy of the system, we show the results by using the Word Accuracy (W_{Acc}), which is defined as $1 - WER$.

It can be observed that the results of the OMR system developed are all close to 90 % of W_{Acc}. This means that a person in charge of the transcription has to deal with just the remaining 10 % to get the perfect transcription of the score, which would result in a very important saving of time and effort.

Fold	T	N_e	C	W_{Acc} (%)
1	390	3	351	93.04
2	377	7	336	89.40
3	623	5	535	87.89
4	432	10	401	90.71
5	410	2	374	92.76
6	427	8	390	90.52
7	514	7	462	90.09
8	436	6	394	90.02
9	441	3	406	92.43
10	444	5	405	91.53
11	633	9	543	87.97
12	641	7	556	90.29
Whole	5768	72	5153	90.36

Table 3 Global results of the OMR systems over the data set. The table contains information about the number of musical symbols in each fold (T), the number of noisy images extracted (N_e) and the number of correct classifications (C). These parameters are used to calculate the Word Accuracy (W_{Acc}).

In order to assess the relevance of our proposal, Table 4 provides a comparison against a previous work that makes use of musical scores from the ACM (see [29]). As mentioned above, the staff detection and removal stage is one of the main reasons for symbol detection losses. The results show that our approach, which circumvents the staff removal process, leads to a remarkably good extraction rate. On the other hand, our classification approach, based on cross-correlation operator, attains good performance.

	Extraction	Classification
Our results	**96.24**	**91.64**
Previous ([29])	72.78	88.86

Table 4 Comparison against previous work with scores from the ACM with average (%) results obtained in the recognition processes.

6 Conclusions

This work presents a new approach to deal with the Optical Music Recognition process for scores written in white mensural notation from the *Archivo de la Catedral de Malaga*. These scores have a special printing style that allows us to propose a new approach in which the very common staff detection and removal stage has been avoided. This stage is critical in the detection and recognition of symbols and it is often one of the main steps to improve the accuracy rates of current OMR systems.

A preprocessing stage is necessary in order to correct both the rotation and the perspective distortion of the input image. At this stage, a binarization process has also been performed to reduce the complexity of the subsequent task. The next stage isolates each staff of the score and a new symbol detection strategy has been followed. This strategy is based on the combination of the use of the *y-projection* of the staff and k-means clustering to detect the boundaries of each symbol region.

These procedures have proven to be reliable as they have achieved extraction rate performance higher than 96 %. The cross-correlation operator has shown its effectiveness in this context for classifying symbols that maintain the staff lines. Classification rates higher than 90 % are attained in all cases. However, new techniques for symbol classification could be applied or developed in future works since there still is some room for improvement. An overall evaluation of the system has also been computed. Our system transcribed the scores with an accuracy close to 90 %.

In comparison with previous results on the ACM (see Table 4), our work attains very good extraction rate, which proves that avoiding staff removal stage is a very valuable choice for the task in terms of symbol detection. In addition, the classification accuracy is also good using a very simple classification strategy.

The work presented opens new avenues for building OMR systems. We believe that the avoidance of the staff detection and removal step deserves further research and can be a way to overcome some of the common misclassification problems that exist in current systems. This approach should be considered to analyze other types of scores to assess if it can be definitely established as a new alternative for the construction of these systems.

References

1. David Bainbridge and Tim Bell. The Challenge of Optical Music Recognition. *Language Resources and Evaluation*, 35:95–121, 2001.
2. I. Barbancho, C. Segura, L.J. Tardon, and A.M. Barbancho. Automatic selection of the region of interest in ancient scores. In *MELECON 2010 - 2010 15th IEEE Mediterranean Electrotechnical Conference*, pages 326–331, April.
3. Ernesto Bribiesca. A new chain code. *Pattern Recognition*, 32(2):235 – 251, 1999.
4. Yung-Sheng Chen, Feng-Sheng Chen and Chin-Hung Teng An optical music recognition system for skew or inverted musical scores. *International Journal of Pattern Recognition and Artificial Intelligence*, 27(07), 2013.
5. Michel M. Deza and Elena Deza. *Encyclopedia of Distances.* Springer, 1 edition, August 2009.
6. Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis.* John Wiley & Sons Inc, 1 edition, February 1973.
7. A. Dutta, U. Pal, A. Fornes, and J. Llados. An efficient staff removal approach from printed musical documents. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 1965–1968, aug. 2010.
8. Alicia Fornés, Josep Lladós, and Gemma Sánchez. Staff and graphical primitive segmentation in old handwritten music scores. In *Proceedings of the 2005 conference on Artificial Intelligence Research and Development*, pages 83–90, Amsterdam, The Netherlands, The Netherlands, 2005. IOS Press.
9. Herbert Freeman. On the encoding of arbitrary geometric configurations. *Electronic Computers, IRE Transactions on*, EC-10(2):260–268, 1961.
10. R. C. Gonzalez and R. E. Woods. *Digital Image Processing.* Prentice-Hall, Upper Saddle River, NJ, USA, 2007.
11. J.A. Hartigan. *Clustering algorithms.* John Wiley & Sons, Inc., 1975.
12. Sun-Kyoo Hwang and Whoi-Yul Kim. Fast and efficient method for computing art. *Image Processing, IEEE Transactions on*, 15(1):112–117, 2006.
13. Frederick Jelinek. *Statistical Methods for Speech Recognition.* The MIT Press, January 1998.
14. VI Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707, 1966.
15. J.P. Lewis. Fast template matching. In *Vision Interface*, pages 120–123, Quebec City, Canada, 1995. Canadian Image Processing and Pattern Recognition Society.
16. K. C. Ng, D. Cooper, E. Stefani, R. D. Boyle, and N. Bailey. Embracing the Composer: Optical Recognition of Handwritten manuscripts. In *Proceedings of the International Computer Music Conference, Beijing*, 1999.
17. N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):62–66, January 1979.
18. J. R. Caldas Pinto, P. Vieira, Mário Ramalho, M. Mengucci, Pedro Pina, and F. Muge. Ancient music recovery for digital libraries. In *Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '00, pages 24–34, London, UK, UK, 2000. Springer-Verlag.

19. João Rogério Caldas Pinto, Pedro Vieira, and João Miguel da Costa Sousa. A new graph-like classification method applied to ancient handwritten musical symbols. *IJDAR*, 6(1):10–22, 2003.
20. S.M. Pizer, R.E. Johnston, J.P. Ericksen, B.C. Yankaskas, and K.E. Muller. Contrast-limited adaptive histogram equalization: speed and effectiveness. In *Visualization in Biomedical Computing, 1990., Proceedings of the First Conference on*, pages 337–345, 1990.
21. D. Pruslin. *Automatic recognition of sheet music*. Sc.d. dissertation, Massachusetts Institute of Technology, 1966.
22. Laurent Pugin. Optical music recognition of early typographic prints using hidden markov models. In *ISMIR*, pages 53–56, 2006.
23. A. Rebelo, I. Fujinaga, F. Paszkiewicz, A.R.S. Marcal, C. Guedes, and J.S. Cardoso. Optical music recognition: state-of-the-art and open issues. *International Journal of Multimedia Information Retrieval*, pages 1–18, 2012.
24. J.N. Sarvaiya, S. Patnaik, and S. Bombaywala. Image registration by template matching using normalized cross-correlation. In *Advances in Computing, Control, Telecommunication Technologies, 2009. ACT '09. International Conference on*, pages 819–822, Dec 2009.
25. M. Sotoodeh and F. Tajeripour. Staff detection and removal using derivation and connected component analysis. In *Artificial Intelligence and Signal Processing (AISP), 2012 16th CSI International Symposium on*, pages 054–057, may 2012.
26. Stephen M. Stigler. Francis Galton's Account of the Invention of Correlation. *Statistical Science*, 4:73–79, 1989.
27. Bolan Su, Shijian Lu, U. Pal, and Chew Lim Tan. An effective staff detection and removal technique for musical documents. In *Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on*, pages 160–164, march 2012.
28. Mariusz Szwoch. A robust detector for distorted music staves. In Andr Gagalowicz and Wilfried Philips, editors, *Computer Analysis of Images and Patterns*, volume 3691 of *Lecture Notes in Computer Science*, pages 701–708. Springer Berlin Heidelberg, 2005.
29. L.J. Tardón, S. Sammartino, I. Barbancho, V. Gómez, and A. Oliver. Optical music recognition for scores written in white mensural notation. *Journal on Image and Video Processing*, 2009:6, 2009.
30. F. Toyama, K. Shoji, and J. Miyamichi. Symbol recognition of printed piano scores with touching symbols. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 2, pages 480–483, 2006.
31. O.D. Trier and T. Taxt. Evaluation of binarization methods for document images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(3):312–315, 1995.
32. Shou-Der Wei and Shang-Hong Lai. Fast template matching based on normalized cross correlation with adaptive multilevel winner update. *Image Processing, IEEE Transactions on*, 17(11):2227–2235, Nov 2008.
33. Charles T. Zahn and Ralph Z. Roskies. Fourier descriptors for plane closed curves. *IEEE Trans. Comput.*, 21(3):269–281, March 1972.