Towards the recognition of compound music notes in handwritten music scores

Arnau Baró, Pau Riba and Alicia Fornés
Computer Vision Center, Dept. of Computer Science
Universitat Autònoma de Barcelona
Bellaterra, Catalonia, Spain
Email: arnau.baro@e-campus.uab.cat, {priba, afornes}@cvc.uab.cat

***Abstract*—The recognition of handwritten music scores still remains an open problem. The existing approaches can only deal with very simple handwritten scores mainly because of the variability in the handwriting style and the variability in the composition of groups of music notes (i.e. compound music notes). In this work we focus on this second problem and propose a method based on perceptual grouping for the recognition of compound music notes. Our method has been tested using several handwritten music scores of the CVC-MUSCIMA database and compared with a commercial Optical Music Recognition (OMR) software. Given that our method is learning-free, the obtained results are promising.**

***Keywords*-Optical Music Recognition; Handwritten Music Scores; Hand-drawn Symbol Recognition, Perceptual Grouping**

I. Introduction

The recognition of music scores [1], [2], [3] has attracted the interest of the research community for decades. Since the first works in the 60s [4] and 70s [5], the recognition or music scores has significantly improved. In the case of printed music scores, one could say that the state of the art has reached a quite mature state. As a matter of fact, many commercial OMR systems show very good performance, such as PhotoScore [1] or SharpEye [2].

Concerning handwritten scores, although it is remarkable the work in Early musical notation [6], [7], the recognition of handwritten Western Musical Notation still remains a challenge. The main two reasons are the following. First, the high variability in the handwriting style increases the difficulties in the recognition of music symbols. Second, the music notation rules for creating compound music notes (i.e. groups of music notes) allow a high variability in appearance that require special attention.

In order to cope with the handwriting style variability when recognizing individual music symbols (e.g. clefs, accidentals, isolated notes), the community has used specific symbol recognition methods [8], [9] and learning-based techniques such as Support Vector Machines, Hidden Markov Models or Artificial Neural Networks [10]. As stated in [11], in the case of the recognition of compound music notes, one must deal not only with the compositional music rules, but also with the ambiguities in the detection and classification of graphical primitives (e.g. note-heads, beams, stems, flags, etc.). It is true that temporal information is undoubtedly helpful in on-line music recognition, as it has been shown in [12], [13]. Nowadays, a musician can find several applications for mobile devices, such as StaffPad[3], MyScript Music[4] or NotateMe[5].

Concerning the off-line recognition of handwritten groups of music notes, much more research is still needed. As far as we know, PhotoScore is the only software able to recognize off-line handwritten music scores, and its performance when recognizing groups of notes is still far from satisfactory. One of the main problems is probably the lack of sufficient training data for learning the high variability in the creation of groups of notes.

For these reasons, in this work we focus on the off-line recognition of handwritten music scores, putting special attention to compound music notes. For this task, we avoid the need of training data and propose a learning-free hierarchical method inspired in perceptual grouping techniques that have been applied to text detection [14] and object recognition [15]. The idea is to hierarchically represent the graphical primitives according to perceptual grouping rules, and then, validate the groupings using music rules.

The rest of the paper is organized as follows. First, the problem statement is described in Section II. Section III describes the preprocessing and the detection of the graphics primitives. Section IV explains the hierarchical representation to combine the graphics primitives into more complex elements, and the validation of each group hypothesis. Section V discusses the experimental results. Finally, conclusions and future work are drawn in Section VI.

II. Problem Statement

Music scores are a particular kind of graphical document that include text and graphics. The graphical information corresponds to staffs, notes, rests, clefs, accidentals, etc., whereas textual information corresponds to dynamics, tempo markings, lyrics, etc.

[1] http://www.neuratron.com/photoscore.htm
[2] http://www.visiv.co.uk/
[3] http://www.staffpad.net/
[4] http://myscript.com/technology/
[5] http://www.neuratron.com/notateme.html

Concerning the recognition of graphical information, Optical Music Recognition (OMR) has many similarities with Optical Character Recognition (OCR). In case of recognizing isolated music symbols (e.g. clefs, accidentals, rests, isolated music notes), the task is similar to the recognition of handwritten characters, digits or symbols. In this sense, the recognizer must deal with the variability in shape, size and visual appearance. Similarly, the recognition of compound music notes (i.e. groups of notes joined using beams) could be seen as the task of recognizing handwritten words.

It is nevertheless true that the difficulties in OMR are higher than in OCR because OMR requires the understanding of two-dimensional relationships, given that music elements are two-dimensional shapes. Indeed, music scores use a particular diagrammatic notation that follow the 2D structural rules defined by music theory. Music notation allows a huge freedom when connecting music notes, which increases the difficulties in the recognition and interpretation of compound notes. For example, music notes can connect horizontally (with beams), and vertically (chords), and the position and appearance highly depends on the pitch (melody), rhythm and the musical effects that the composer has in mind. Figure 1 shows several examples of compound music groups that are equivalent in rhythm.

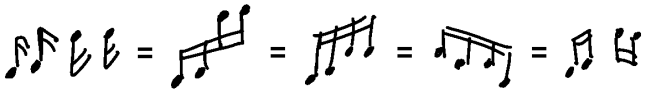


Figure 1: Equivalent (in rhythm) compound Sixteenth notes.

Following the comparison with handwritten text recognition, it is true that language models can also be defined to improve the OMR results, just like language models and dictionaries help in handwriting recognition. For example, syntactical rules and grammars could be easily defined to cope with the ambiguities in the rhythm. In music theory, the time signature defines the amount of beats per bar unit. Therefore, all the music notes inside a bar unit must sum up to the defined amount of beats. Although grammars and rules [16], [17] have shown to be very useful to solve ambiguities, it is extremely difficult to use them when there are several melodic voices and chords, such as in polyphonic music. Moreover, it must be said that music *tuplets* (defined as irrational rhythms or extra-metric groupings) and ornament notes (e.g. *Appoggiatura*) scape from the beats restriction.

Finally, semantics could also be defined using knowledge modeling techniques (e.g. ontologies). Indeed, a musicologist could define the harmonic rules that should be applied for dealing with melodic ambiguities in polyphonic scores. However, these rules highly depend on the composer and the time period (e.g. some dissonant chords or intervals are only common in modern ages). Therefore, the incorporation of this knowledge seems unfeasible in this OMR stage.

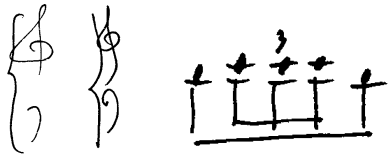
III. Preprocessing and Detection of Primitives

In the preprocessing, we remove music braces and ties. In this step we assume that the input image is binary and the staff lines are already removed by using any of the staff removal methods in the literature [18]. Then we detect the graphics primitives: note-heads, vertical lines and beams.

A. Preprocessing

1) Brace removal: In polyphonic scores, braces indicate the staffs that are played together, such as scores for different instruments. Given that braces appear at the beginning, we analyze the connected components at the beginning of the staffs. Following the musical notation theory, a brace must cross consecutive staffs. Thus, these elements are approximated to a straight line, and if the estimated line crosses several staffs, it is classified as a brace. Afterwards, braces are removed using the straight line estimation in order to avoid the deletion of other elements such as clefs. The removal of braces will ease the posterior recognition of music symbols, such as clefs and key-signatures. Figure 2a) shows some examples of braces where some of them are overlapping the clefs. For more details, see [19].

2) Tie removal: Long ties are used for adding expressibility in music performance. However, they can be easily misclassified as beams due to the handwriting style of the musician. Figure 2b) shows a problematic case, where the beam is disconnected from the stems. Therefore, we propose to detect and remove the long ties by analyzing the aspect ratio of the horizontally long connected components.



(a) (b)

Figure 2: a) Examples of braces that are gathered with clefs. b) Beam easily confused as a tie.

B. Detection of Graphics Primitives

The starting point to construct the proposed hierarchical representation is the detection of basic primitives that defines the musical vocabulary or compound notes. These basic primitives are created by means of simple detectors.

1) Vertical lines detection: Vertical lines are key elements that are mainly used to represent stems and bar lines. Since music notes are mainly composed by note-heads, stems, beams and flags (see Fig.3), we must identify the bar lines so that we can keep the rest of vertical lines as stem candidates. For this task, we first detect all the vertical lines using a median filter, and then, we analyze them to identify the bar lines. The bar line identification consists of two steps:

- *Properties checking*: The vertical line is kept as a bar line candidate if it (almost) crosses all the staff and it has no blobs (note-heads) at it extrema points.
- *Consistency checking*: The bar lines in the same page must have similar length and must cross the same staffs. Therefore, the consistency is analyzed as follows. First, we vertically sort the bar line candidates using their centroid. Here, one candidate is an outlier if its length is very different from the candidates in the same line. These outliers are analyzed just in case they have not been correctly detected so they must be joined with other vertical lines. Otherwise, they are rejected.

2) Note-head detection: Note-heads play a key-role in music notes, since they provide the melody. Moreover, its the only common component in all type of music notes. Hence, detecting correctly a note-head is of key importance for the correct symbol construction. Figure 3 shows in red the different types of note heads that must be detected.

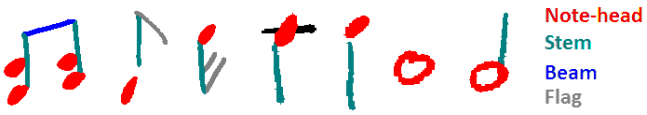


Figure 3: Graphics Primitives.

Filled-in note-heads are detected using mathematical morphology. First, two elliptic structural elements are defined using different angles (30^o and -30^o). Then a morphological closing is performed using both structural elements. Finally, blobs closer to a vertical line are considered filled-in note-heads. For the detection of white note-heads, the filled-in note-heads are first removed from the image. Then, the holes are filled so that we can find white note-heads using the same strategy. In both cases, too large blobs are rejected.

3) Beam detection: The beam's appearance highly depends on the melody. Consequently, a descriptor based on densities, profiles or gradients (e.g. SIFT, HOG) will be unstable. For this reason, we propose the detection of beams by adapting a pseudo-structural descriptor [20] for handwritten word spotting. The feature vector is created from the information from every key-point in the word. For each key-point, the characteristic Loci Features encode the frequency of intersection counts following a certain direction path. Thus, the shape of the strokes is not taken into account.

For the detection of beams, we propose to modify the pseudo-structural descriptor as follows. For each pair of consecutive detected note-heads (and stems), we take the region in between, and divide it into 2 parts (left and right). Then, we compute the characteristic Loci Features in the vertical direction (i.e. the number of transitions). Finally, we take the statistical mode (the most frequent value), which indicates the amount of beams that link each pair of notes (see Fig. 4). In this way, the descriptor is invariant to the beam's appearance and orientation.

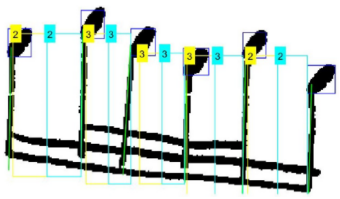


Figure 4: Detected primitives in a compound note. The numbers indicate the number of detected beams per region.

IV. Perceptual Grouping

Once we have detected the graphics primitives, the next step consists in grouping them to recognize the compound music notes. First, we create a hierarchical representation of primitives (see Fig.5), and then we validate the different grouping hypothesis using syntactical rules.

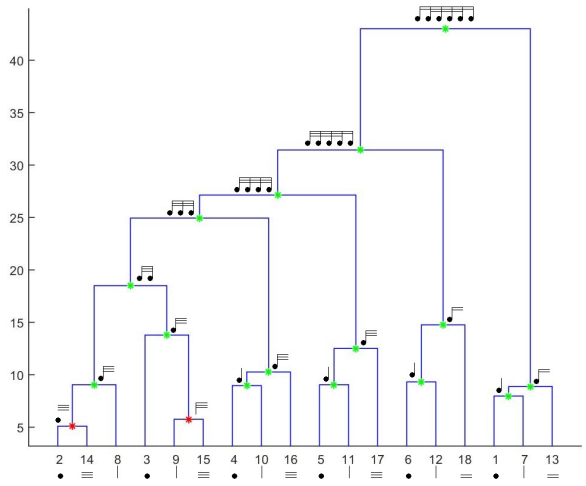


Figure 5: Validation hypothesis (dendrogram) of the compound note shown in Figure 4.

A. Hierarchical representation

Inspired in the perceptual grouping techniques for text detection [14] and object recognition [15], we propose to build a dendrogram to hierarchically represent the graphics primitives. In our case the criteria for grouping is the proximity of the graphics primitives, which means that the coordinates of the primitives' centers are used as features to create the hierarchical clustering.

Since compound music notes must contain at least one note-head, we use the detected note-head candidates as seeds to start the grouping in a bottom-up manner. Thus, we avoid the creation of many non-meaningful grouping regions.

Notice that the different grouping hypothesis can overlap. For instance, a chord is composed of several note-heads that share the same stem (e.g. see the first note in Fig. 3). Thus, this stem belongs to more than one group hypothesis.

B. Validation of grouping hypothesis

The next step is the validation of the groupings. In case of text detection, the grouping validation could be performed by recognizing the text. For example, a grouping hypothesis could be accepted whenever an OCR can recognize the word.

In our case, the recognition of the compound notes as a whole is not possible because the creation of a dictionary of music notes is unfeasible: there are almost-infinite combinations of compound notes. Moreover, we would need an huge amount of samples to train a shape recognizer. Therefore, we propose to validate each one of the grouping hypothesis through the following music notation rules:

- Whole note = {[white-note-head]+ }.
- Half note = {[white-note-head]+, stem}.
- Quarter note = {[filled-in-note-head]+, stem}.
- 8th note = {[filled-in-note-head]+, stem, beam}.
- 16th note = {[filled-in-note-head]+, stem, beam, beam}.
- 32th note = {[filled-in-note-head]+, stem, beam, beam, beam}.
- 64th note = {[filled-in-note-head]+, stem, beam, beam, beam, beam}.

The symbol + indicates that minimum one appearance of this primitive is required. In summary, only the grouping hypothesis that can be validated using these rules will remain. All the other hypothesis will be rejected.

V. Evaluation

For the experiments, we have selected a subset of the CVC-MUSCIMA dataset [21]. Concretely, we have manually created the ground-truth of 10 music pages, which contain a total of 1932 music notes. The music scores are from 4 different writers, mostly polyphonic music (containing several voices and chords). As stated in the introduction, since we focus on the recognition of compound music notes, we leave out of our experiments the recognition of isolated symbols (e.g. clefs, accidentals), which could be faced with symbol recognition methods, as shown in [8].

Table I shows the experimental results. The first column indicates the music page that has been used (e.g. 'w5-p02' means page 2 from writer 5). The second column indicates whether the score is polyphonic or monophonic. The third and forth columns show the detection of note-heads, whereas the last two columns show the detection of music notes (e.g. half, quarter, 8th note, etc.). The metrics used are the Precision (number of correctly detected elements divided by the number of detected elements), and Recall (number of correctly detected elements divided by the number of elements in the dataset).

We observe that the mean Precision and Recall of music notes is around 52%. The main reason is that the detection of note-heads (which are used as seeds in the grouping) is sensitive to the handwriting style. For example, in scores from writer 10, the head-note detector misses almost half of the note-heads. Consequently, the detection of music notes is always lower that this value. In some other cases, such as scores from writers 17 and 38, the note-head detector works much better, which in turn allows the music notes detector to be much higher (recall is 68% and 78%, respectively).

Table I: Results. The detection of note-heads and music notes are shown in terms of Precision (P) and Recall (R). All results are between [0-1].

Score	Polyphonic	Note-heads		Notes	
		P	R	P	R
w5-002	No	0,6	0,62	0,49	0,5
w5-010	Yes	0,63	0,62	0,36	0,35
w5-011	No	0,58	0,6	0,48	0,5
w5-012	Yes	0,72	0,73	0,64	0,65
w10-002	No	0,61	0,67	0,47	0,52
w10-010	Yes	0,62	0,61	0,4	0,39
w10-011	No	0,64	0,54	0,55	0,47
w10-012	Yes	0,59	0,55	0,49	0,45
w17-012	Yes	0,64	0,73	0,6	0,68
w38-012	Yes	0,76	0,82	0,72	0,78
Mean	-	0,64	0,65	0,52	0,53

Our method has been compared with PhotoScore[6], a commercial OMR software able to recognize handwritten music scores. Figures 6 and 7 show qualitative results from both approaches. As it can be noticed, PhotoScore performs very well in easy parts, whereas its performance decreases considerably in case of complex compound music notes. In this aspect, our approach is much more stable.

Table II: Comparison with the commercial PhotoScore OMR software. Detection of music notes in terms of Precision (P) and Recall (R). All results are between [0-1].

Score	PhotoScore		Our method	
	P	R	P	R
w10-010	0,63	0,61	0,4	0,39
w38-012	0,69	0,74	0,72	0,78

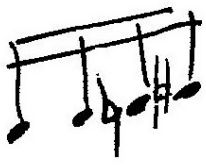


Figure 8: Compound notes with accidentals.

Table II shows the quantitative results. As it can be seen, our method outperforms the recognition of compound music notes 'w38-012'. Contrary, the differences in the recognition of the 'w10-010' score are very high. There are two main

[6]http://www.neuratron.com/photoscore.htm

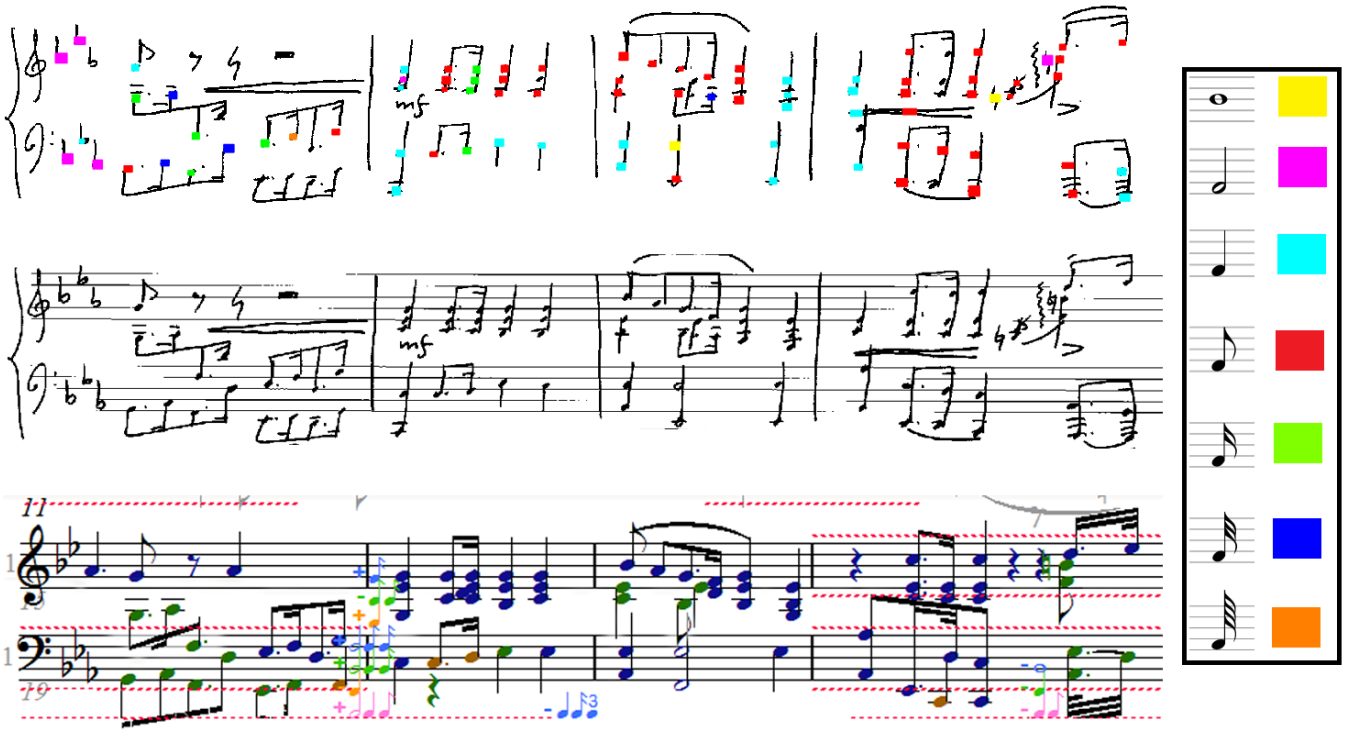


Figure 6: Results on ‘w10-p10’. First row: our method. Second row: original image. Third row: PhotoScore results.

reasons: first, our limitation of correctly detecting note-heads (the recall is around 60% in this score); and secondly, the accidentals (e.g. sharps or naturals) that appear inside the compound music symbols (see Fig.8) create confusion in the dendrogram. In addition, flats are similar to half notes, and they are frequently confused.

In any case, it must be said that this comparison is not completely fair. PhotoScore has some features to improve its performance that are not considered in our method. First, PhotoScore is a complete OMR system that recognizes the whole score, which probably uses training data to deal with the variability in the handwriting style. Since it recognizes all music symbols (including clefs, accidentals and rests), it can use syntactic rules for validation. For instance, the system can recognize the time signature and then validate the amount of music notes at each bar unit (which is used to solve ambiguities).

VI. Conclusion

In this work we have proposed a learning-free method for recognizing compound groups of music notes in handwritten music scores. Our method is composed of a hierarchical representation of graphics primitives, perceptual grouping rules and a validation strategy based on music notation. Since our method does not use any training data, the experimental results are encouraging, especially when compared with a commercial OMR software.

As a future work, we would like to improve the detection of note-heads because it is clearly limiting the performance of our method. In this sense, a more sophisticated key-point detector for note-heads should be investigated. Moreover, we also plan to recognize isolated symbols by using symbol recognition so that we can incorporate syntactical rules (e.g. time measure checking). Finally, we plan to test our method with scores from many more writers.

Acknowledgment

This work has been partially supported by the Spanish project TIN2015-70924-C2-2-R and the *Ramon y Cajal* Fellowship RYC-2014-16831. The authors thank Lluis Gomez for his suggestions on perceptual grouping.

References

[1] D. Bainbridge and T. Bell, “The challenge of optical music recognition,” *Computers and the Humanities*, vol. 35, no. 2, pp. 95–121, 2001.

[2] A. Rebelo, I. Fujinaga, F. Paszkiewicz, A. Marcal, C. Guedes, and J. Cardoso, “Optical music recognition: state-of-the-art and open issues,” *IJMIR*, vol. 1, no. 3, pp. 173–190, 2012.

[3] A. Fornés and G. Sánchez, “Analysis and recognition of music scores,” in *Handbook of Document Image Processing and Recognition*. Springer-Verlag London, 2014, pp. 749–774.

[4] D. Pruslin, “Automatic recognition of sheet music, phd thesis,” Massachusetts, USA, 1966.

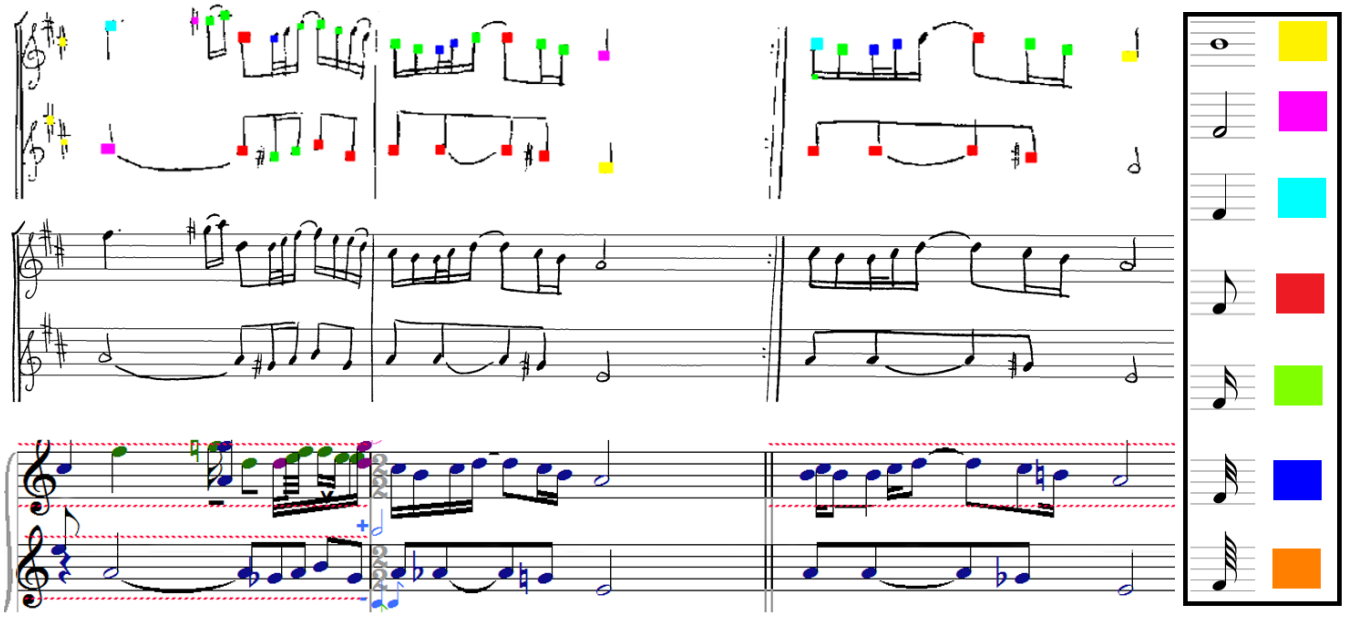


Figure 7: Results on'w38-p012'. First row: our method. Second row: original image. Third row: PhotoScore results.

[5] D. Prerau, "Computer pattern recognition of standard engraved music notation, phd thesis," Massachusetts, USA, 1970.

[6] J. C. Pinto, P. Vieira, and J. M. Sousa, "A new graph-like classification method applied to ancient handwritten musical symbols," *IJDAR*, vol. 6, no. 1, pp. 10–22, 2003.

[7] L. Pugin, "Optical music recognition of early typographic prints using hidden markov models," in *International Conference on Music Information Retrieval*, 2006, pp. 53–56.

[8] A. Fornés, J. Lladós, G. Sánchez, and D. Karatzas, "Rotation invariant hand drawn symbol recognition based on a dynamic time warping model," *IJDAR*, vol. 13, no. 3, pp. 229–241, 2010.

[9] S. Escalera, A. Fornés, O. Pujol, P. Radeva, G. Sánchez, and J. Lladós, "Blurred Shape Model for binary and grey-level symbol recognition," *Pattern Recognition Letters*, vol. 30, no. 15, pp. 1424–1433, 2009.

[10] A. Rebelo, G. Capela, and J. S. Cardoso, "Optical recognition of music symbols: A comparative study," *IJDAR*, vol. 13, no. 1, pp. 19–31, 2010.

[11] K. Ng, "Music manuscript tracing," in *Graphics Recognition Algorithms and Applications*. Springer, 2001, pp. 330–342.

[12] H. Miyao and M. Maruyama, "An online handwritten music symbol recognition system," *IJDAR*, vol. 9, no. 1, pp. 49–58, 2007.

[13] J. Calvo-Zaragoza, J.; Oncina, "Recognition of pen-based music notation with probabilistic machines," in *7th International Workshop on Machine Learning and Music*, Barcelona, Spain, 2014.

[14] L. Gomez and D. Karatzas, "Multi-script text extraction from natural scenes," in *12th ICDAR*. IEEE, 2013, pp. 467–471.

[15] N. Ahuja and S. Todorovic, "From region based image representation to object discovery and recognition," in *Structural, Syntactic, and Statistical Pattern Recognition*. Springer, 2010, pp. 1–19.

[16] H. Kato and S. Inokuchi, *Structured Document Image Analysis*. Springer-Verlag, 1991, ch. A recognition system for printed piano music using musical knowledge and constraints, pp. 435–455.

[17] S. Macé, É. Anquetil, and B. Coüasnon, "A generic method to design pen-based systems for structured document composition: Development of a musical score editor," in *1st Workshop on Improving and Assessing Pen-Based Input Techniques*, 2005, pp. 15–22.

[18] M. Visani, V. C. Kieu, A. Fornés, and N. Journet, "ICDAR 2013 music scores competition: Staff removal," in *12th ICDAR*. IEEE, 2013, pp. 1407–1411.

[19] P. Riba, A. Fornés, and J. Lladós, "Towards the alignment of handwritten music scores," in *Eleventh IAPR International Workshop on Graphics Recognition (GREC)*, 2015.

[20] J. Lladós, M. Rusinol, A. Fornés, D. Fernández, and A. Dutta, "On the influence of word representations for handwritten word spotting in historical documents," *International journal of pattern recognition and artificial intelligence*, vol. 26, no. 05, p. 1263002, 2012.

[21] A. Fornés, A. Dutta, A. Gordo, and J. Lladós, "CVC-MUSCIMA: a ground truth of handwritten music score images for writer identification and staff removal," *IJDAR*, vol. 15, no. 3, pp. 243–251, 2012.