Global constraints for syntactic consistency in OMR: an ongoing approach

Ana Rebelo[1], André Marçal [2], Jaime S. Cardoso[1]

[1]INESC TEC, Faculdade de Engenharia, Universidade do Porto, Portugal
[2]CICGE, Faculdade de Ciências, Universidade do Porto, Portugal
arebelo@inescporto.pt, andre.marcal@fc.up.pt, jaime.cardoso@inescporto.pt

Abstract. Optical Music Recognition (OMR) systems are an indispensable tool to transform the paper-based music scores and manuscripts into a machine-readable symbolic format. A system like this potentiates search, retrieval and analysis. One of the problematic stages is the musical symbols detection where operations to localize and to isolate musical objects are developed. The complexity is caused by printing and digitalization, as well as the paper degradation over time. Distortions inherent in staff lines, broken, connected and overlapping symbols, differences in sizes and shapes, noise, and zones of high density of symbols is even worst when we are dealing with handwritten music scores. In this paper the exploration of an optimization approach to support semantic and syntactic consistency after the music symbols extraction phase is proposed. The inclusion of this ongoing technique can lead to better results and encourage further experiences in the field of handwritten music scores recognition.

Key words: Computer Vision, Image Processing, Optical Music Recognition

1 Introduction

Prior to music typographical systems, all music was copied manually including large scores and each and every part for the players and singers. There remains a substantial and important corpus of works that exist only as original hand-written manuscripts or facsimiles of these manuscripts. These important cultural artifacts are in danger of being lost through the normal ravages of time. To preserve the music (rather than the documents themselves) an optical music recognition (OMR) system is needed. Notwithstanding, for the several techniques already existent in the literature the results in the handwritten symbol segmentation are still far from ideal. One way to improve this can be achieved by exploring higher order knowledge presented globally for the whole music sheet. Being these documents well structured pieces of work, in this paper a method to incorporate syntactic and semantic music rules after the segmentation and recognition process in order to overcome possible errors is presented. These rules are described as an optimization problem encompassing global constraints about musical rules.

The segmentation and classification process has been the object of study in the research community with several proposed techniques (e.g. [1–4]). The introduction of the musical context in this process has been suggested through grammars or graphs (e.g. [5–9, 1]).

The most usual segmentation approach decomposes the music sheet hierarchically. First, the score is analyzed and split by staffs and then the primitive symbols are extracted (e.g. [4]). Some authors make the segmentation step along with the classification operation (e.g. [3]) and others prefer to separate them (e.g. [10, 9]).

Rossant and Bloch [3] proposed to use region growing, template matching and Hough Transform to extract the music symbols. A fuzzy model with recognition hypotheses is used to include contextual information and music writing rules. In [11] the symbols are extracted using a connected components process and small elements are removed based on their size and position on the score. One of Fujinaga's first works focused on the characterization of music notation by means of a *context-free* and $LL(k)$ grammar [5]. Coüasnon [8] also formalizes the musical knowledge by a grammar, while Carter [6] uses a Line Adjacency Graph (LAG) to extract symbols. A structural method based on the construction of graphs for each symbol is proposed by Randriamahefa [7]. Bainbridge [9] also implemented a grammar-based approach to specify the relationships between the recognized musical shapes.

For the classification of the symbols most of the authors (e.g. [12, 13, 4]) suggested the extraction of some features. Choudhury et al. [12] proposed the extraction of width, height, area, number of holes and low-order central moments of the detected objects, whereas Taubman [13] extracted standard moments, centralized moments, normalized moments and Hu moments. The classifiers adopted were the k-nearest neighbor (kNN), the Mahalanobis distance and the Fisher discriminant. An investigation on four classification methods, namely Support Vector Machines (SVMs), Neural Networks (NNs), Nearest Neighbour (kNN) and Hidden Markov Models, was also carried out in [4].

Most of the methods use complex processes in an already complex problem to include syntactic and semantic rules. We believe that using content knowledge of the image and music writing rules it is possible to simplify the procedure and also obtain good results. Reducing the processing time of execution could also lead to other applications such as online music recognition, taking advantage of new small electronic devices with increasing computation power.

After the introduction where a brief description of the work done in the OMR field was made, this paper is organized as follows: in Section 2 we present the syntactic consistency model, detailing the optimization problem in Subsection 2.1. The experimental results are described in Section 3 and finally the main conclusions are outlined in Section 4.

2 Syntactic Consistency

During the detection and classification of the music symbols errors always occur: missed, missclassified and falsely detected symbols. The purpose of the introduction of syntactic and semantic musical rules before the construction of the final musical notation model is to overcome these possible errors. In this work, the last two problems were treated.

The main idea behind the syntactic consistency procedure is related to the fact that in music sheet the number of symbols contained within two bar lines and the time signature must match – see Fig. 1. This type of sign is sometimes represented by two figures where the numerator occupies the two top spaces and the denominator occupies the two bottom spaces, or simply it could be represented, for instance, by a 𝄴 used to indicate $\frac{4}{4}$ or a 𝄵 used to indicate $\frac{2}{2}$. The procedure of checking the coherency is entirely related to the *measure*. The top number in the time signature indicates the number of beats to be counted in each measure/bar line, while the bottom number indicates which type of note value equals one beat. For instance, the number of beats or the number of symbols between barlines in the Fig. 1 is 2 and the symbol that indicates the unit is the minim symbol (𝅗𝅥), because the bottom number is 2, giving $\chi \times 2 = 1 \Leftrightarrow \chi = 1/2$, where χ represents the symbol type. Every measure of music in a simple time signature has the same number of beats per measure throughout the song. Hence, symbols confusion and wrong symbols added can be mitigated by querying if the detected symbols' durations match with the amount of the value of the time signature on each bar.

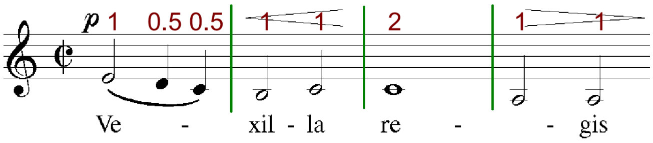

Fig. 1. Example of measures according to the time signature. The bar lines are represented by the green lines.

2.1 Global Constraints

We propose to detect the best combination of symbols between two bar lines given the indicated measure in the music sheet as an optimization problem. In this manner, the syntactic and semantic music rules can be incorporated as global constraints, considering the following:

$$
\begin{array}{ll}
\max & \displaystyle\sum_{i=1}^{n}\sum_{j=1}^{k} p_{ij}x_{ij} \\
 & \sum_{j=1}^{k} x_{ij} \leq 1, i = 1, \ldots, n \\
 & D\sum_{j=1}^{k}\sum_{i=1}^{n} \alpha_j x_{ij} = N \\
s.t & \sum_{i=1}^{n} x_{i2} \leq \sum_{i=1}^{n}\sum_{j=3}^{M} x_{ij} \\
 & x_{i,1} \leq \sum_{j=3}^{M} x_{i+1j}, i = 1, \ldots, n-1 \\
 & x_{ij} \in \{0, 1\}
\end{array}
\qquad (1)
$$

where p_{ij} is the likelihood (matrix of probabilities) given by a classifier, of the symbol i to belong to the class j, x_{ij} represents the symbol i from class j, N and D are the numerator and denominator (as aforementioned) in the time signature, respectively, n is the total number of symbols, k is the total number of classes, M is the total number of symbols that could have associated accidentals and accents, and α_j is the music note value that represents how long each note lasts. Since notes come in different levels, each with its own note value, it is possible to associate a note value to each class j. The first constraint of the optimization problem allows symbols elimination, the second constraint is related to the time signature, the third and fourth constraints are related to the position of the symbols in the staff.

Accents are placed above beams, below and above noteheads. Accidentals exist before each notehead and at same height (placed on the same staff line or space). The third constraint incorporates the rule of accents position on the optimization problem, allowing accents only if notes are presented in the score. The fourth constraint imposes precedence of the accidentals for the note symbols.

To better understand how α_j works in the second constraint see Fig. 2 which illustrates most of the notes that we can find in music arranged. For instance, the value of a half note (𝅗𝅥) is half of a whole note (𝅝), the value of a quarter note (𝅘𝅥) is quarter of a whole note (𝅝), and so on. In Fig. 1 the note value of 𝅗𝅥is 1/2. The rests have the same values as the notes. Occasionally, in a music sheet we can have dots placed to the right of notes and rests, increasing the original value: n dots lengthen the note's or rest's original d duration to $d \times (2 - 2^{-n})$. These α_j values will be used in the counting of the time between the barlines. The note value of accents and accidentals is 0, because they do not interfere in the meter. Depending on the time signature, the number of beats per note varies. Nevertheless, the note lengths do not change in their relationships to each other: one beat of music could indicate the length of a whole note, but two quarter notes will always be twice as fast as a half note.

The matrix of probabilities given by p_{ij} is the result of a classifier. In this work a Multi-Layer Perception (MLP) classifier with one hidden layer with an hyperbolic tangent sigmoid activation function was used. The input of the network was the symbol bounding box resized to 20 × 20 pixels converted to a vector of binary values with more 23 features: the percentage of black pixels, the orientation of the symbol, the number of vertical and horizontal holes, the ratio

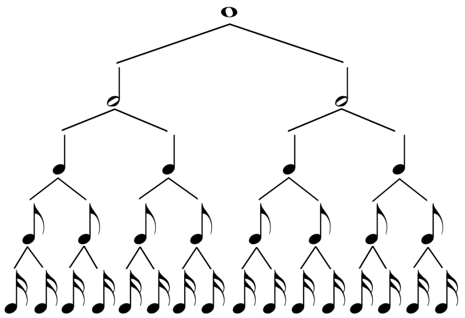

Fig. 2. Tree of notes representing the relation between them. At the top is the whole note, below that half notes, then quarter notes, eighth notes, and finally sixteenth notes.

between volume and connected components area, the number of end points and intersections in the object skeleton and features from the Blurred Shape Model (BSM) descriptor (see [14] for more details). These features were based on the Gamera project[1] and follow common practices from the state-of-the-art methods in the OMR field [15].

Additional experiments were conducted with a majority vote combination of three MLP classifiers named Combined Neural Networks (CNN). In here two of the networks have the same architecture, but the initial random weights are different. The third network is fed with a different input and with a different number of neurons in the hidden layer. For two of the networks each image of a symbol was initially resized to 20 × 20 pixels and then converted to a vector of 400 binary values. For the third network each image of a symbol was initially resized to 60 × 20 pixels and then converted to a vector of 1200 binary values. Usually the images have an height larger than their width and the idea was thus to favor the height. In this manner, the problem in the classification of barlines, due to its similarity with dots after the resize, is minimized. Like in the first case more 23 features were added to the raw pixels. Once again, all these networks have one hidden layer with an hyperbolic tangent sigmoid transfer function.

3 Evaluation Metrics and Results

The data set adopted to test the proposed classifiers and the syntactic consistency process consists of 6 handwritten scores from 3 different composers, 9 synthetic scores and 9 scanned printed scores, all written on the standard notation. The scanned and real scores were binarized using Otsu's method [16], while the synthetic scores were already in binary format. In total we work with 1713 handwritten music symbols, 6486 printed music symbols and 4267 synthetic music symbols.

[1] http://gamera.informatik.hsnr.de

Two databases, one for printed (synthetic plus scanned) music symbols and another one for handwritten music symbols, of training patterns were created to obtain the models to classify. The choice of working with two databases instead of one is related to the high variability of the handwritten objects that we wanted to avoid in the results of the printed scores. Each one of these databases was randomly split into training and test sets, with 25% and 75% of the data, respectively. This division was repeated 50 times with a training set with classes proportionally represented. The best parametrization of each model was found using the training and validation sets, with the expected error estimated on the test set by a 3-cross validation scheme. The results for the different models can be seen in Table 1. We see a better accuracy with printed scores and also when we increase the number of features, as expected. The high variability of the symbols in real scores imposes no difference between a single MLP and three combined MLPs with 423 features.

	MLP + 400	MLP + 423	CNN + 423
Printed scores	[87%; 89%]	[88%; 90%]	[95%; 96%]
Real scores	[81%; 82%]	[82%; 83%]	[82%; 83%]

Table 1. 99% CI for the expected performance for the classification models using 400 features and 423 features.

To test the performance of our syntactic consistency model we used the reference position of the music symbols on the music score. This reference position was obtained manually and it is composed by the coordinates of the bounding box of the object plus its class. The aim was to verify if the optimization process improves the classification results, making a good re-label of the objects without the segmentation interference. Table 2 presents the obtained results where TPB and TPA stands for true positive classification before and after the execution of the proposed syntactic consistency process, respectively.

	MLP + 400		MLP + 423		CNN + 423	
	TPB	TPA	TPB	TPA	TPB	TPA
Synthetic scores	83%	82%	82%	83%	94%	92%
Real scores	74%	73%	74%	73%	84%	82%
Scanned scores	78%	79%	79%	79%	79%	81%

Table 2. Accuracy obtained for the classification models using 400 features and 423 features before and after the syntactic consistency model.

In some situations, with printed and scanned scores, we could improve the overall performance. It is clear that the adjustment of the parameters of the global optimization problem is necessary. Nevertheless, the final accuracy loss was not significant to stop the work on the global constraints methodology. Changing the constraints, making possible to include new symbols or weight the re-label of the symbols could improve the performance.

4 Conclusion

The work developed in this paper had the target to overcome the several issues that affect musical symbol recognition especially in handwritten scores. We proposed a parametric model to incorporate syntactic and semantic music rules after a music symbols segmentation's method. The optimization problem detects the best combination of symbols based on global constraints that ensures the indicated measure. The results obtained make us to continue to believe that including prior knowledge in the OMR recognition process it leads to better results.

Acknowledgments This work is financed by National Funds through the FCT Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project SFRH/BD/60359/2009.

References

1. Bellini, P., Bruno, I., Nesi, P.: Optical music recognition: Architecture and algorithms. In: Interactive Multimedia Music Technologies. Hershey: IGI Global (2008) 80–110
2. Forns, A., Llads, J., Snchez, G.: Primitive segmentation in old handwritten music scores. In Liu, W., Llads, J., eds.: GREC. Volume 3926 of Lecture Notes in Computer Science., Springer (2005) 279–290
3. Rossant, F., Bloch, I.: Robust and adaptive OMR system including fuzzy modeling, fusion of musical rules, and possible error detection. EURASIP Journal on Advances in Signal Processing **2007**(1) (2007) 160–160
4. Rebelo, A., Capela, G., Cardoso, J.S.: Optical recognition of music symbols: A comparative study. International Journal on Document Analysis and Recognition **13** (2009) 19–31
5. Prerau, D.: Optical music recognition using projections (1988) In D. Blostein and H. Baird. A Critical Survey of Music Image Analysis. In *Structured Document Image Analysis*, pages 405–434, Springer-Verlag, Heidelberg, 1992.
6. Carter, N.P.: Automatic recognition of printed music in the context of electronic publishing (1989) In Dorothea Blostein and Henry S. Baird, *A Critical Survey of Music Image Analysis*, in *Structured Document Image Analysis*, Baird, Bunke, and Yamamoto (Eds.), Eds., pp. 405–434, Springer-Verlag, Heidelberg, 1992.
7. Randriamahefa, R., Cocquerez, J., Fluhr, C., Pepin, F., Philipp, S.: Printed music recognition. Proceedings of the Second International Conference on Document Analysis and Recognition (Oct 1993) 898–901
8. Coüasnon, B.: Segmentation et reconnaissance de documents guides par la connaissance a priori: application aux partitions musicales. PhD thesis, Universit de Rennes (1996)

9. Bainbridge, D.: An extensible optical music recognition system. Nineteenth Australasian Computer Science Conference (1997) 308–317
10. Fujinaga, I.: Staff detection and removal. In George, S., ed.: Visual Perception of Music Notation: On-Line and Off-Line Recognition. Idea Group Inc. (2004) 1–39
11. Tardón, L.J., Sammartino, S., Barbancho, I., Gómez, V., Oliver, A.: Optical music recognition for scores written in white mensural notation. EURASIP Journal on Image and Video Processing **2009** (February 2009) 6:3–6:3
12. Choudhury, G.S., Droetboom, M., DiLauro, T., Fujinaga, I., Harrington, B.: Optical music recognition system within a large-scale digitization project. In: International Society for Music Information Retrieval (ISMIR). (2000)
13. Taubman, G., Odest, A., Jenkins, C.: Musichand: A handwritten music recognition system. Technical report (2005)
14. Escalera, S., Fornés, A., Pujol, O., Radeva, P., Sánchez, G., Lladós, J.: Blurred shape model for binary and grey-level symbol recognition. Pattern Recognition Letters **30** (November 2009) 1424–1433
15. Rebelo, A., Tkaczuk, J., Sousa, R., Cardoso, J.S.: Metric Learning for Music Symbol Recognition. In: The tenth International Conference on Machine Learning and Applications (ICMLA'11). (2011)
16. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man and Cybernetics **9**(1) (1979) 62–66