Goal-Directed Evaluation for the Improvement of Optical Music Recognition on Early Music Prints

Laurent Pugin, John Ashley Burgoyne, and Ichiro Fujinaga
Schulich School of Music, McGill University
555 Sherbrooke West
Montreal, QC H3A 1E3
{laurent,ashley,ich}@music.mcgill.ca

ABSTRACT

Optical music recognition (OMR) systems are promising tools for the creation of searchable digital music libraries. Using an adaptive OMR system for early music prints based on hidden Markov models, we leverage an edit-distance evaluation metric to improve recognition accuracy. Baseline results are computed with new labeled training and test sets drawn from a diverse group of prints. We present two experiments based on this evaluation technique. The first resulted in a significant improvement to the feature extraction function for these images. The second is a goal-directed comparison of several popular adaptive binarization algorithms, which are often evaluated only subjectively. Accuracy increased by as much as 55% for some pages, and the experiments suggest several avenues for further research.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Information Storage and Retrieval—*performance evaluation (efficiency and effectiveness)*; I.7.5 [**Document and Text Processing**]: Document Capture—*document analysis*; J.5 [**Arts and Humanities**]: *performing arts*

General Terms

Algorithms, Performance, Reliability, Experimentation

Keywords

Optical music recognition. Early music. Adaptive binarization. Goal-directed evaluation. Test-driven development.

1. INTRODUCTION

Optical music recognition (OMR) tools can be used to create searchable digital libraries from previously inaccessible musical content, a feature especially useful for historical music sources [6]. Using OMR on this scale, however, is a complex task requiring a large infrastructure [2], and as with optical character recognition (OCR) tools, many parameters must be taken into account, most notably the accuracy of the OMR tool to be used. Assessing the accuracy of OMR systems remains an open question [1]. All current commercial tools act as "black boxes": only their final, high-level results are available to researchers. OMR tools that make low-level symbolic information available, such as Gamera [5], are easier to evaluate for accuracy.

This paper presents an example of taking advantage of an evaluation metric to improve an OMR system for an early-music digitization project with images presenting degradation. We use Aruspix, an adaptive OMR tool for early typographic prints using hidden Markov models (HMMs) [7], which evaluates accuracy as the edit distance between low-level symbol sequences. We tested Aruspix on various pages taken from prints in several different music fonts. After analyzing the results, we modified the feature extraction method used in Aruspix. Several adaptive binarization algorithms were able to improve performance further. In the absence of a robust testing infrastructure, these algorithms are often evaluated only subjectively [4]; the Aruspix infrastructure enabled us to use objective, goal-directed evaluation [9], to our knowledge the first use of this approach in OMR.

2. EVALUATION AND EXPERIMENTS

Early music has been a challenge for traditional approaches to OMR [6]; HMMs, however, the approach taken in Aruspix, have proven to be very effective [7]. As HMMs are based on learning, good training and test sets are necessary to optimise their performance.

2.1 Baseline

To obtain a broader training set than was available in [7], we increased the ground-truth data by transcribing prints produced with eight different music fonts by printers from Italy (3), France (3), Belgium (1) and Germany (1). A new typographic model was built based on the 457 available pages. The number of symbols in the set of pages is 220 (vs. 175 in [7]) and the number of characters 95,845 (vs. 52,170 in [7]). Because so many aspects of these prints (music font used, degree of document degradation, scanning settings, etc.) are unpredictable, all of variability in the data cannot reasonably be represented in a single training set. This limits the value of cross-validation. To better evaluate what the results would be in a real-world application, we created a test set of 40 pages taken from eight other prints produced with eight other music fonts. These prints were from Italy (6), France (1) and Belgium (1). Unlike any

JCDL'07, June 18–23, 2007, Vancouver, British Columbia, Canada.

of the prints in the training set, two of these used unusual printing techniques (Antico was produced with woodblocks and Petrucci in multiple impressions), which enabled us to evaluate Aruspix when pushed beyond its original intended usage.

2.2 Experiments with feature extraction

When retrained and tested on the new test set, the typographic model showed a lack of robustness to changes in staff-line width, which can vary considerably due to printing irregularities. In response, we modified the feature extraction (FE) function presented in [7]. We measure the line width while detecting staff curvature, which precedes FE. This measured width is then subtracted when calculating 3 of the 7 features in the original function: the total area, the largest black element, and the center of vertical gravity.

2.3 Experiments with AB

Bleed-through (elements from the other side of the page that are visible through the paper, e.g., Figure 1) affected several prints in the test set. This type of degradation is a problem for many historical documents, musical and non-musical. In some case, bleed-through can be reduced during binarization. Numerous solutions have been proposed, including adaptive binarization (AB), which examines a small window around each pixel in an image before making the foregound-background decision [8]. We tried four different AB algorithms with window sizes of 3, 7, 15, 31, 63 and 127: Bernsen (B) [8], Gatos (G) [3], Niblack (N) [8] and Sauvola (S) [8]. The results were compared with the global binarization method of Otsu [8] used in Aruspix previously. In order to be able to evaluate the effect of window size consistently, we applied AB after deskewing and resizing the image to normalize all staves to a height of 100 pixels.

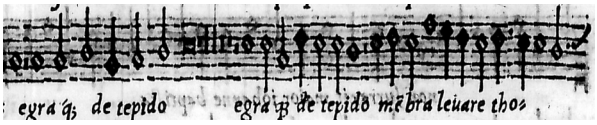

Figure 1: Example with bleed-through from Buglhat

3. RESULTS

Table 1 shows the average recognition rates for each print in the test set: the baseline, the results after improvement with the FE method, the results with AB using S-15 (the algorithm and window which give the best average results over the whole set), and the results with the best AB algorithm for the specific print (determined *a posteriori*). The results confirm that the FE phase is crucial to the process: without including any new features, we were able to increase the accuracy significantly. Only with the Waelrant print did the recognition rate decrease with the FE modification, which may be because the strokes in this print are finer than the others in the set. The results also show that although AB can be very helpful, different AB algorithms perform best for different prints. Moreover, a logistic regression model showed that for the prints on which AB helps significantly, it is very sensitive to region size, with the ideal size being around 15 pixels, or slightly less than the size of a staff space. Where AB does not help significantly, it can in fact reduce accuracy when the region size is less than 31 pixels. The regression model also showed a highly significant interaction between the prints and the algorithms, and even between individual pages and the algorithms.

Table 1: Recognition rates

Print	Base.	FE	AB (S-15)	With best AB
Petrucci	75.91	79.22	79.56	80.59 (B-31)
Antico	35.14	56.03	78.38	79.94 (N-7)
Moderne	69.09	81.03	88.75	88.75
Buglhat	54.01	69.89	84.29	84.29
Waelrant	94.76	92.87	93.12	93.12
Rampazetto	86.95	90.52	80.80	90.52 (S-127)
Sabbio	59.49	76.75	87.38	89.67 (G-3)
Bessozzi	80.15	87.11	79.42	88.46 (S-127)
Whole set	69.34	79.18	84.23	86.85

4. CONCLUSION AND FUTURE WORK

Evaluating Aruspix against a new test set highlighted potential improvements to its recognition system. Because Aruspix is adaptive and can learn from labeled data, we were able to increase accuracy rapidly, 35% to 80% for Antico, without modifying the training or test sets nor using cross-validation on a fused training and test set. The great variability in early music sources makes such a test-driven approach particularly effective. Goal-directed evaluation of several AB algorithms also revealed that comparing binarization algorithms is too subtle a task for subjective human experts.

How to choose the ideal algorithm and window size, however, remains an open question. It will be very important to identify features of these prints that can automatically predict the ideal binarization technique. With more labeled training data, which one might reasonably expect to have in the context of a large-scale digitization project, we would also expect further improvements in accuracy and greater robustness to the wide variability of early music prints.

5. ACKNOWLEDGMENTS

We would like to thank the Canada Foundation for Innovation and the SSHRC for their support of this project.

6. REFERENCES

[1] P. Bellini, I. Bruno, and P. Nesi. Assessing optical music recognition tools. *Computer Music Journal*, (forthcoming).

[2] G. S. Choudhury, T. DiLauro, M. Droettboom, I. Fujinaga, B. Harrington, and K. MacMillan. Optical music recognition system within a large-scale digitization project. In *Proc. 1st Int. Conf. on Music Inf. Ret.*, 2000. On-line presentation.

[3] B. Gatos, I. Pratikakis, and S. J. Perantonis. An adaptive binarisation technique for low quality historical documents. In *Proc. IAPR Workshop on Document Analysis Systems*, number 3163 in LNCS, pages 102–13, 2004.

[4] E. Kavallieratou and E. Stamatatos. Adaptive binarization of historical document images. In *Proc. 18th ICPR*, pages 742–745, 2006.

[5] K. MacMillan, M. Droettboom, and I. Fujinaga. Gamera: Optical music recognition in a new shell. In *Proc. Int. Computer Music Conf.*, pages 482–485, 2002.

[6] J. C. Pinto, P. Vieira, M. Ramalho, M. Mengucci, P. Pina, and F. Muge. Ancient music recovery for digital libraries. In *Proc. 4th ECDL*, pages 24–34, 2000.

[7] L. Pugin. Optical music recognition of early typographic prints using hidden markov models. In *Proc. 7th Int. Conf. on Music Inf. Ret.*, pages 53–56, 2006.

[8] M. Sezgin and B. Sankur. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging*, 13(1):146–65, 2004.

[9] O. D. Trier and A. K. Jain. Goal-directed evaluation of binarization methods. *IEEE Transactions on PAMI*, 17(12):1191–1201, 1995.