Incremental learning for recognition of handwritten mensural notation

Luisa Micó [1] **José M. Iñesta** [1] **David Rizo** [1]

Abstract

This paper presents an ongoing research on handwritten symbol recognition in early music scores. The help of human supervision is needed for a correct edition and publication of these collections. A suitable strategy is needed for optimizing the exploitation of human feedback to improve and adapt the classifier to the specificities of each manuscript. The objective is to minimize the number of interactions needed to solve the problem, thus optimizing the user workload. The strategy is shown to be convenient but there is still work ahead for improving its performance.

1. Introduction

This work is in the context of a project on collections of handwritten scores of Spanish music from the 16th and 17th centuries, coded in the Spanish white mensural notation system. We are working with a collection of vocal polyphonic music, although the scores are separated parts of individual voices, so the images are of monophonic staves. See Fig. 1 for an example.

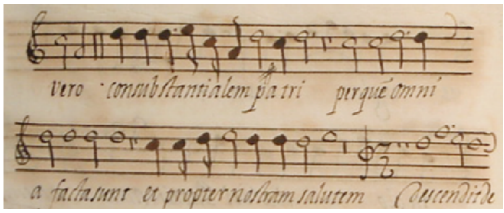
vero consubſtantialem pa tri per quē omni
a facta sunt et propter noſtram salutem descendit de

Figure 1. The kind of documents processed (here a fragment of a page) are in Spanish white mensural notation.

The automatic pattern recognition approach has been traditionally focused on accomplishing a fully-automated operation. State-of-the-art optical music recognition (OMR) systems are still far from a perfect performance, much less when dealing with handwritten documents. The management of the errors produced by the system is usually seen as an issue outside the research process because it is simply considered as the procedure for converting the system hypothesis into the desired result. However, semi-automatic approaches in which the human operator has the eventual responsibility of verifying and completing the task are gaining importance in the field (Toselli et al., 2011).

A perfect transcription of the original documents is needed for editing and publishing a collection. Therefore we have to focus on the human-machine interaction tasks and how to optimize the expert user feedback loop. The automatic transcription of this kind of music documents into a symbolic format is a complex task (Rebelo et al., 2012), and therefore it is adequate to consider interactive paradigms (Sober-Mira et al., 2017).

In this paper we will show how using user's corrections helps the classifier to learn its model incrementally, lowering the error rate along the task.

2. Method

The recognition algorithm is a key issue in any pattern classification system, but in an interactive architecture, the most important feature is the ability of the algorithm to adapt to the data specificities through the error corrections made by the user (Oncina, 2009). In the interactive paradigm, the efficient exploitation of *human expert knowledge* is the main objective, so the correctness of the system output is no longer the main issue to assess. The challenge now is the development of interactive schemes capable of efficiently exploiting the human feedback in order to eventually reduce the user's workload.

In light of that, we have selected a very simple, but flexible, classification algorithm as the nearest neighbor is. It does not need a parametric analysis of the feature space for operation, and the training set $\mathcal{X}$ can be incrementally built by adding new pairs as they are found in the input in operation time: $\mathcal{X}^{(k+1)} = \mathcal{X}^{(k)} \bigcup \{\mathbf{x}_i, y_i\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^d$ is a new feature vector and $y_i \in \mathcal{C}$ is the label associated to that sample. Even it is easy to add new classes dynamically by adding new labels, if needed. Also, editing and condensing

[1] Dept.of Software and Computing Systems, University of Alicante, Spain. Correspondence to: José M. Iñesta <inesta@dlsi.ua.es>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

methods (Wilson & Martinez, 2000) can be easily applied to the training set if the corrections made by the user demand those operations to improve the performance. The system must operate in real time, so the user can interact with it comfortably. This is another feature that advices to use simple, adaptive, and fast classification algorithms.

2.1. System setup

This work is based on the simulation of a user correcting the hypothesis made by the system. In real operation, a collection of scores is presented to the user by pages. Each page is processed and the symbols are classified (see (Calvo-Zaragoza, 2016) for details). Then, the user makes corrections to the symbols that were incorrectly classified either due to misclassification or because a symbol belongs to a previously unseen class. These interactions are utilized to improve the model for the classification of the next pages.

The initial model is trained with the symbols in the first page: $\mathcal{X}^{(1)} = \mathcal{X}^{(P_1)}$, including the labels for the symbols in it, $y_i \in \mathcal{C}^{(P_1)}$. The feature vectors for the symbols consist of gray level values for the pixels in the region of interest (ROI) where each symbol is found. These regions are normalized to a 30×30 pixel square ROI, so $\mathbf{x}_i \in [0, 255]^{30\times30}$.

The algorithm outline is shown in Alg. 1. As explained, errors in $\mathcal{E}^{(p)}$ can be due to symbols belonging to unseen classes. In such case, the interaction step includes the addition of the new class to the training set, with the symbols seen in the current page as prototypes. This algorithm will try to minimize the number of errors $\ell^{(p)}$, and therefore the need of user corrections, as the pages are being processed.

Algorithm 1 Outline of the method

```
Input: Collection of pages $\mathcal{P} = \{P_1, P_2, ..., P_{|\mathcal{P}|}\}$
$\mathcal{X}^{(1)} = \mathcal{X}^{(P_1)}$
Train the model $\mathcal{M}^{(1)}$ with $\mathcal{X}^{(1)}$
for $p = 2$ to $|\mathcal{P}|$ do
    Apply $\mathcal{M}^{(p-1)}$ to samples in $\mathcal{P}_p$
    $\mathcal{E}^{(p)} = \{\mathbf{x}_i \in \mathcal{P}_p \ / \ \hat{y}_i \neq y_i\}$
    $\ell^{(p)} = |\mathcal{E}^{(p)}|$
    Interaction: the user fixes wrong $\hat{y}_i$ to actual $y_i$
    $\mathcal{X}^{(p)} = \mathcal{X}^{(p-1)} \bigcup \mathcal{X}^{(p)}$
    Train $\mathcal{M}^{(p)}$ with $\mathcal{X}^{(p)}$
end for
```

3. Data and results

As a proof of concept, we have selected one of the collections we are studying, composed of 57 pages of a Mass in A minor. Each page has 6 staves, containing between 20 and 30 symbols in average, including notes, rests, and other symbols. From the symbols in P_1, 19 different classes were established that were increased until 25 at the end of the collection (see Fig. 2). A rapid drop of the error was initially observed to less than a half of the original percentage. The error rises again when a new composition begins to be processed, due to a change in the clef that causes a change in the relative position of many notes, so their stems are in the opposite direction. When this effect is learned by the model, it is able to lower the error again. Every time a change in the conditions happens the system degrades its performance, but it is able to continue learning later on.

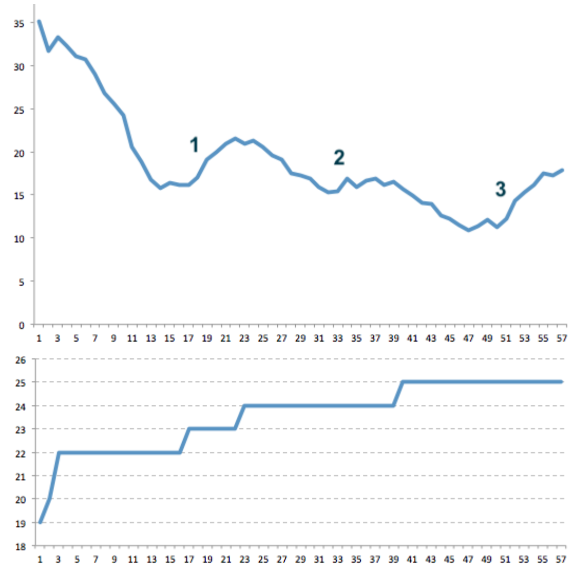

Figure 2. (Top): Evolution of a smoothed version of the error rate (%) along the pages in a particular collection composed of 57 pages. Mark ‘1’ corresponds to a change to a new composition written in a different clef (from C to G). Mark ’2’ locates a new composition written again in C clef. In mark ’3’ the way of digitizing the scores were changed making a big different in how they look like. (Bottom): Evolution of the number of classes in the training set, aligned to the pages in the top graph.

4. Conclusion

The presence of the expert user in the learning loop opens up new possibilities for study adaptive learning algorithms. The presented study has to be regarded as a initial proof of concept and these results have to be analyzed from the qualitative point of view rather than quantitatively. In any case, they are promising enough to encourage us to keep on searching for new classification methods and ways to make the human-in-the-loop approach efficient.

Acknowledgements

Work supported by the Spanish Ministry HISPAMUS project TIN2017-86576-R, partially funded by the EU.

References

Calvo-Zaragoza, J.; Rizo, D.; Iñesta J.M. Two (note) heads are better than one: pen-based multimodal interaction with music scores. In Devaney, J. et al. (ed.), *17th International Society for Music Information Retrieval Conference*, pp. 509–514, New York City, August 2016. ISBN 978-0-692-75506-8.

Oncina, Jose. Optimum algorithm to minimize human interactions in sequential computer assisted pattern recognition. *Pattern Recognition Letters*, 30:558–563, 2009.

Rebelo, Ana, Fujinaga, Ichiro, Paszkiewicz, Filipe, Marçal, André R. S., Guedes, Carlos, and Cardoso, Jaime S. Optical music recognition: state-of-the-art and open issues. *International Journal of Multimedia Information Retrieval*, 1(3):173–190, 2012.

Sober-Mira, Javier, Calvo-Zaragoza, Jorge, Rizo, David, and Iñesta, José Manuel. Pen-based music document transcription. In *Proceedings of the 12th IAPR International Workshop on Graphics Recognition, GREC 2017, Kyoto, Japan, November 9-10, 2017*, 2017.

Toselli, Alejandro Hctor, Vidal, Enrique, and Casacuberta, Francisco. *Multimodal Interactive Pattern Recognition and Applications*. Springer Publishing Company, Incorporated, 1st edition, 2011. ISBN 0857294784, 9780857294784.

Wilson, D. Randall and Martinez, Tony R. Reduction techniques for instance-based learning algorithms. *Machine Learning*, 38(3):257–286, 2000.