2017 14th IAPR International Conference on Document Analysis and Recognition

Pen-based Music Document Transcription

Javier Sober-Mira,[1] Jorge Calvo-Zaragoza,[2] David Rizo,[1] and José M. Iñesta [1]
[1]Software and Computing Systems, University of Alicante, Alicante, Spain
[2]Schulich School of Music, McGill University, Montréal, Canada
jsober@dlsi.ua.es, jorge.calvozaragoza@mail.mcgill.ca, {drizo,inesta}@dlsi.ua.es

***Abstract*—The transcription of music sources requires new ways of interacting with musical documents. Assuming that automatic technologies will never guarantee a perfect transcription, our intention is to develop an interactive system in which user and software collaborate to complete the task. Since the use of traditional software for score edition might be tedious, our work studies the interaction by means of electronic pen (e-pen). In our framework, users trace symbols using an e-pen over a digital surface, which provides both the underlying image (offline data) and the drawing made (online data). Using both sources, the system is capable of reaching an error below 4% when recognizing the symbols with a Convolutional Neural Network.**

***Index Terms*—Music Documents; Optical Music Recognition; Pen-based technologies; Convolutional Neural Networks;**

I. Introduction

Automatic recognition systems have been traditionally focused on accomplishing a fully-automated operation, yet optimum performance cannot be assured. 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. Quite often, however, we find a semi-automatic scenario in which the human operator has the eventual responsibility of verifying and completing the task [1].

In this paper, we focus on the human-machine interaction for tasks related to music notation. The automatic transcription of music documents into a symbolic format is a complex task [2], and so it seems suitable to consider interactive paradigms when dealing with it. However, conventional channels of communication such as the keyboard or the mouse are not easily applicable. However, handwriting is a natural way of communication for humans and therefore it is interesting to consider it for interacting with the computer. This can be done by means of electronic pen (e-pen) technologies.

Although the user is provided with a friendlier interface to interact, interaction is no longer deterministic: unlike the keyboard or mouse entry, for which it is clear what the user is inputting, the pen-based interaction has to be decoded and this process might have errors [3].

This work focuses on the use of pen-based technologies for the human-computer interaction involving music documents. This task produces an interesting multimodal signal, which can be used to boost the recognition. To carry out this classification task, we make use of Convolutional Neural Networks. Within this paradigm, we can nicely combine the different modalities produced so that the performance can be improved as far as possible.

II. Materials and methods

We assume a workflow in which the user traces symbols on a digital surface depicting a music score. The system, therefore, receives a multimodal signal: on one hand, the sequence of points that indicates the path followed by the e-pen on the digital surface —usually referred to as *online* modality; on the other hand, the piece of image below the drawn, which contains the original traced symbol —*offline* modality. The goal is to use this interaction to transcribe the musical document. Since the interaction itself gives us the position of the symbols in the image, it is only necessary to infer which type of symbol has been traced in each interaction.

Figure 1 illustrates the process explained above for a single symbol. The actual information obtained is the sequence of 2D points in the same order they were collected, indicating the path followed by the e-pen (*online* information). An image representation of the symbol can be rendered by generating segments between pairs of consecutive points. In addition, we can consider the bounding box of the pen strokes to crop the original image, thereby obtaining the symbol of interest as it appears in the original image (*offline* information).

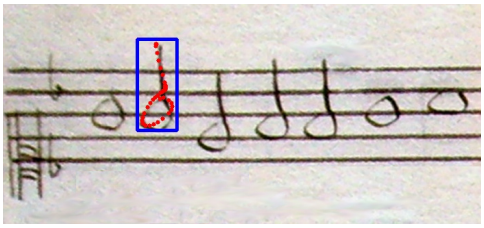
To carry out our experiments, we consider a dataset consisting of 10 150 samples, each of which is represented by both offline and online modalities. Data was collected by five different users tracing symbols on an archive of early music handwritten in Mensural notation. These samples are spread over 30 classes. The number of symbols per class is not balanced but it depicts the same distribution found in the documents.

III. Convolutional Neural Network

We base our classification on Convolutional Neural Networks (CNN), given their great success in a range of tasks related to computer vision [4]. These networks take advantage of local filters, pooling, and many connected layers to learn a suitable data representation for classification tasks.

We denote by $Conv(k, c)$ a spatial convolutional layer with kernel size $k \times k$ and number of filters c, with Rectifier Linear Unit activation. Similarly, we denote by $MaxPool(k)$ a max-pooling layer with kernel size $k \times k$. $Dropout(r)$ is a dropout procedure with a ratio of dropped units r. As a proof of concept, we have evaluated our approach with the following configuration: $Conv(32, 3) \rightarrow Conv(32, 3) \rightarrow MaxPool(2) \rightarrow Conv(32, 3) \rightarrow Conv(32, 3) \rightarrow MaxPool(2)$.

2379-2140/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDAR.2017.258



(a) Tracing process

(b) Offline data

(c) Online data

Fig. 1. Example of extraction of a *minima*. Above, the sequence of points collected by the e-pen. The box represents the bounding box of the sequence. Below, the multimodal data extracted.

A fully connected layer with 30 units and *SoftMax* activation is also placed on top of the CNN in order to obtain a probability for each of the considered categories.

IV. Classification

All images (from both online and offline modalities) are rescaled to a fixed size of 36×36 pixels before being provided to the CNN.

Moreover, several strategies can be considered to address the classification of the multimodal data:

1) Single modality: It is interesting to measure how well the single modalities considered behave by themselves. To this end, we consider the *single mode* classification strategy, which means that only one modality is used for the classification.

2) Late fusion: Due to the *SoftMax* layer, the output of the CNN corresponds to values between 0 and 1, indicating the confidence that the network gives to each possible category. Therefore, decisions of independent CNN can be merged by a linear combination.

Let Ω denote the set of categories considered. Given images x and y from the offline and online modality, respectively, the late fusion emits the label $\hat{\omega}$ such that

$$\hat{\omega} = \arg\max_{\omega \in \Omega} \frac{1}{2} P_{\text{off}}(\omega|x) + \frac{1}{2} P_{\text{on}}(\omega|y),$$

where $P_{\text{off}}(\omega|x)$ and $P_{\text{on}}(\omega|y)$ are the probabilities obtained from the corresponding modality.

V. Results

Experimentation followed a 5-fold cross-validation scheme. The independent folds were randomly created with the sole constraint of having the same number of samples per class (when possible) in each of them.

TABLE I
Average results obtained for a 5-fold cross validation experiments with respect to the classification scheme.

Classification strategy	Error rate $\pm$ std. dev.
Offline modality	6.3 ± 0.7
Online modality	7.3 ± 0.2
Late fusion	3.6 ± 0.5

Table I shows the error rate achieved by each combination of network model and classification scheme. When considered isolatedly, the offline modality achieves better results than the online one. This is probably because the handwriting style of the users with the e-pen is higher than that depicted in the music documents.

In any case, the experiments report that the multimodal classification (late fusion) significantly outperforms the strategies that only use a single modality. In this case, less than $4\ \%$ of the symbols are mislabeled. Therefore, the proposed approach is presented as an interesting tool for transcribing musical documents.

VI. Conclusions

This paper presents a new approach to transcribe music documents into a computer by using e-pen technologies. Our framework produces a multimodal signal with which to improve the music symbol classification.

Experimentation with a particular corpus was presented, considering CNN and several classification schemes. Results support that it is worth to consider both modalities in the classification process, as accuracy is noticeably improved with a combination of them than that achieved by single modalities.

This is a first step to achieve a complete system for music document transcription. More factors are still of interest, such as detecting the position of the symbols in the staff.

Acknowledgment

This work was supported by the Social Sciences and Humanities Research Council of Canada, and by the Spanish Ministerio de Economía y Competitividad through Project TIMuL (No. TIN2013-48152-C2-1-R supported by EU FEDER funds).

References

[1] A. H. Toselli, E. Vidal, and F. Casacuberta, *Multimodal Interactive Pattern Recognition and Applications*, 1st ed. Springer Publishing Company, Incorporated, 2011.

[2] A. Rebelo, I. Fujinaga, F. Paszkiewicz, A. R. S. Marçal, C. Guedes, and J. S. Cardoso, "Optical music recognition: state-of-the-art and open issues," *International Journal of Multimedia Information Retrieval*, vol. 1, no. 3, pp. 173–190, 2012.

[3] J. Calvo-Zaragoza, D. Rizo, and J. M. I. Quereda, "Two (note) heads are better than one: Pen-based multimodal interaction with music scores," in *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016, New York City, United States, August 7-11, 2016*, 2016, pp. 509–514.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–44, 2015.