Multimodal Recognition for Music Document Transcription

Javier Sober-Mira, Jorge Calvo-Zaragoza, David Rizo, and José M. Iñesta

Department of Software and Computing Systems
University of Alicante, Alicante, Spain
{jsober,jcalvo,drizo,inesta}@dlsi.ua.es

Abstract. Converting sheet music scores into symbolic format is a necessary step to use computational tools for music indexing and analysis. We present an interactive framework in which user and computer collaborate to complete this transcription task. Our scenario assumes that the user traces the information found in the image using an electronic pen, and the system automatically recognizes the music symbols. The multimodal nature of the signal (both pen strokes and source image) can be used to improve the automatic recognition with Convolutional Neural Networks. Our experiments show that exploiting adequately this multimodality leads to a lower error rate.

1 Introduction

A large number of sheet music sources are available for musicological study. An interesting option is to use computational tools for large-scale indexing and analysis of the music. However, for this task to be feasible it is necessary to have that content transcribed into a machine-readable format.

An efficient way of digitizing sheet music is to resort to automatic transcription tools, usually referred to as Optical Music Recognition. These systems try to automatically extract the meaningful information contained in a music document from an image of its source. Nevertheless, these systems are still far from solving the problem accurately [4], which finally makes this option be discarded in most of the cases. Then, manual transcription is the only option left.

It is important to emphasize the role of the user as part of this process. In such case, the user is the most valuable resource and the system must be focused on minimizing the effort needed to complete the task [5]. It is, therefore, necessary to develop tools that allow an intuitive and efficient interface. In spite of several efforts to develop light and friendly software for music score edition, the process is still considered tedious by most users.

We focus on the human-machine interaction for tasks related to music document transcription. Conventional channels such as keyboard or mouse are not easily applicable here, and so there is a need to introduce new ways of interaction. Handwriting is a natural way of communication for humans, and so it would be interesting to use this kind of interaction for music document transcription. This can be done by means of electronic pen (e-pen) technologies.

The scenario stated produces multimodal signals that can be used to improve the recognition of the music symbols. In this work, we extend the first step presented in [1] by considering Convolutional Neural Networks for the multimodal classification. Within this paradigm, several ways of combining the different modalities produced are proposed, so that the performance can be improved as far as possible.

2 Multimodal data

This section describes the nature of the multimodal data considered in this work. The corpus of our case of study consists of 60 scores from a music archive dated between centuries 16th to 18th, handwritten in mensural notation [2].

We assume a framework in which the user traces symbols on the score using a digital surface, with the aim of automatically recognizing the music symbols. The system therefore receives a multimodal signal: on the one hand, the piece of the original image below the traced shape, referred to as *offline* modality (Fig. 1); on the other hand, the rendered image of the sequence of 2D points followed by the e-pen on the surface, referred to as *online* modality (Fig. 1). The challenge here is how to achieve an adequate synergy that eventually allows taking maximum advantage of all the modalities involved.

Fig. 1: Offline modality

Fig. 2: Online modality

The considered dataset consists 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 the aforementioned archive.

The samples are spread over 30 classes. The number of symbols of each class is not balanced but it depicts the same distribution found in the documents.

2.1 Data overview

The distribution of symbols in the data set is shown in Fig. 3. In each cross-validation folder, a similar number of representatives from each class is found. For example, there are about 2600 samples of the *Minima* symbol, so in each cross-validation set there are approximately 500 samples of it, and the same applies to all classes. In any case, the maximum difference among the sizes of the cross-validation folders for any class was of 21 samples.

Some of the most important symbols considered are shown in Table 1.

Group	Symbol			
Note	Semibrevis	Minima	Col. Minima	Semiminima
				
Rest	Longa	Brevis	Semibrevis	Semiminima
				
Clef	C Clef	G Clef	F Clef (I)	F Clef (II)
				
Signature	Major	Minor	Common	Cut
				
Others	Flat	Sharp	Dot	Custos
				

Table 1: A representative subset of the elementary symbols of the mensural notation archive considered in this work.

Also it is important to point out that those who helped to create the on-line mode data were writers with a basic knowledge of music, not specialists in early music notations. They were not instructed in how the specific mensural symbols should be interpreted.

The online sequences created by the different writers were shuffled in a single set and the cross-validation folders were taken randomly from it, in such a way that the different folds were not conditioned by a particular writing style.

3 Classification framework

We base our classification on Convolutional Neural Networks (CNN), given their great success in a range of tasks related to computer vision [3]. These networks take advantage of local filters, pooling, and many connected layers for learning a data representation to successfully solve classification tasks.

The topology of these networks can be very varied. We selected four architectures that are described below. 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. Then, our network architectures are defined sequentially as:

1. Conv(32,3) → Conv(32, 3) → MaxPool(2) → Conv(32, 3) → Conv(32, 3) → MaxPool(2)
2. Conv(32, 3) → Conv (32, 3) → MaxPool(2) → Conv(32, 3) → Conv(32, 3) → MaxPool(2) → Dropout(0.1)

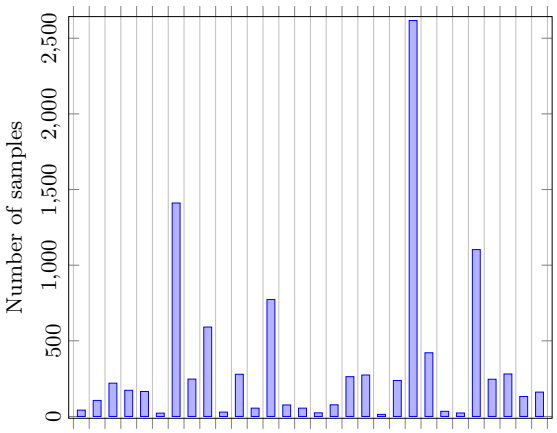

Fig. 3: Distribution of the 10 150 samples among the 30 different classes considered. The most repeated symbols are *Minima* (2617), *Col. Minima* (1411), *Semibrevis* (1103), *Col. Semiminima* (591), and *Dot* (773).

3. Conv(64, 3) → Conv (64, 3) → MaxPool(2) → Conv(32, 3) → Conv(32, 3) → MaxPool(2) → Conv(16, 3) → MaxPool(2) → Conv(16, 3) → MaxPool(2)
4. Conv(64, 3) → Conv (64, 3) → MaxPool(2) → Conv(32, 3) → Conv(32, 3) → MaxPool(2) → Conv(16, 3) → MaxPool(2) → Conv(16, 3) → MaxPool(2) → Dropout(0.1)

In all cases, a fully connected layer with 30 units and *SoftMax* activation is placed on top of the CNN in order to obtain a probability for each of the considered categories. Also, the input layers always expect images of an equal size of 36×36.

3.1 Multimodal classification

There might be several multimodal classification strategies depending on where the modality fusion is actually performed. Next lines describe each of the strategies considered.

Single mode It is interesting to consider how well the single modalities considered behave in order to assess the goodness of the fusion modalities. To this end, we consider the *single mode* classification strategy, which means that just a single modality is considered for the classification.

Late fusion *Late fusion* tries to merge the classification decisions obtained for each modality in order to obtain a more robust decision that takes into account both sources of information at the same time.

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 networks 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, this fusion emits the label $\hat{\omega}$ such that

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

Where $P_{\text{off}}(\omega|x)$ and $P_{\text{on}}(\omega|y)$ are the probabilities given by the network used for the corresponding modality. Note that α is a parameter that tunes the weight given to each single modality. This parameter has to be fixed empirically.

Intermediate fusion Taking into account the internal operation of the CNN, an *intermediate fusion* can be considered. That is, the combination is performed in the intermediate layers of the network.

In this case, the intermediate union is achieved by the concatenation of the CNN used for each modality. Afterwards, a new fully connected layer with *SoftMax* activation is added to output a new probability value for each category. A graphical illustration is given in Fig. 4,

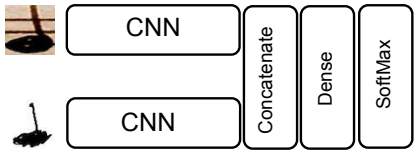

Fig. 4: Structure of the whole multimodal-model classification scheme.

4 Experimentation

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 2 shows the error rate achieved by each combination of network model and classification scheme. Several $\alpha \in [0, 1]$ were tested for the late fusion strategy, and the best results were obtained for $\alpha = 0.5$.

Our experiments report that the information fusion (both late and intermediate) behave better than using single modalities, satisfying our initial hypothesis.

	Model 1	Model 2	Modal 3	Modal 4
Single mode (offline)	6.3 ± 0.7	6.1 ± 0.7	7.6 ± 0.2	6.6 ± 1.0
Single mode (online)	7.3 ± 0.2	7.3 ± 0.3	10 ± 2.1	7.7 ± 0.8
Late fusion ($\alpha = 0.5$)	3.6 ± 0.5	**3.2 ± 0.2**	4.3 ± 0.8	4.4 ± 0.4
Intermediate fusion	3.5 ± 0.7	3.5 ± 0.6	4.2 ± 0.8	3.6 ± 0.6

Table 2: Error rate (average ± std. deviation) obtained for a 5-fold cross validation experiments with respect to the classification scheme and CNN model.

The best results, on average, are reported by the late fusion with the network model 2. However, the difference with other models does not seem to be really significant.

5 Conclusion

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

Results with this particular dataset was presented, considering CNN and several multimodal 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 each single modality.

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. For this task, an initial approach is using other CNN with images that have more top and bottom margins, so that it is able to discriminate by position. Those margins should be enough to see the whole staff, and thus be able to detect the vertical position in the same way a human could do.

References

1. Jorge Calvo-Zaragoza, David Rizo, and José Manuel Iñesta 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*, pages 509–514, 2016.
2. Antonio Ezquerro Esteban. Música de la catedral de barcelona a la biblioteca de catalunya. *Biblioteca de Catalunya, Barcelona*, 2001.
3. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
4. Ana Rebelo, Ichiro Fujinaga, Filipe Paszkiewicz, André R. S. Marçal, Carlos Guedes, and Jaime S. Cardoso. Optical music recognition: state-of-the-art and open issues. *International Journal of Multimedia Information Retrieval*, 1(3):173–190, 2012.
5. Alejandro H. Toselli, Verónica Romero, Moisés Pastor, and Enrique Vidal. Multimodal interactive transcription of text images. *Pattern Recogn.*, 43(5):1814–1825, May 2010.