# TWO (NOTE) HEADS ARE BETTER THAN ONE: PEN-BASED MULTIMODAL INTERACTION WITH MUSIC SCORES

**Jorge Calvo-Zaragoza, David Rizo, Jose M. Iñesta**
Pattern Recognition and Artificial Intelligence Group
Department of Software and Computing Systems
University of Alicante, Spain
{jcalvo,drizo,inesta}@dlsi.ua.es

## ABSTRACT

Digitizing early music sources requires new ways of dealing with musical documents. Assuming that current technologies cannot guarantee a perfect automatic transcription, our intention is to develop an interactive system in which user and software collaborate to complete the task. Since conventional score post-editing might be tedious, the user is allowed to interact using an electronic pen. Although this provides a more ergonomic interface, this interaction must be decoded as well. In our framework, the user traces the symbols using the electronic pen over a digital surface, which provides both the underlying image (offline data) and the drawing made by the e-pen (online data) to improve classification. Applying this methodology over 70 scores of the target musical archive, a dataset of 10 230 bimodal samples of 30 different symbols was obtained and made available for research purposes. This paper presents experimental results on classification over this dataset, in which symbols are recognized by combining the two modalities. This combination of modes has demonstrated its good performance, decreasing the error rate of using each modality separately and achieving an almost error-free performance.

## 1. INTRODUCTION

Music constitutes one of the main tools for cultural transmission. That is why musical documents have been preserved over the centuries, scattered through cathedrals, museums, or historical archives. In an effort to prevent their deterioration, access to these sources is not always possible. This implies that an important part of this historical heritage remains inaccessible for musicological study. Occasionally, these documents are transcribed to a digital format for easier access, distribution and study, without compromising their integrity.

On the other hand, it is important to point out that the massive digitization of music documents also opens several opportunities to apply Music Information Retrieval algorithms, which may be of great interest. Since the manual transcription of these sources is a long, tedious task, the development of automatic transcription systems for early music documents is gaining importance in the last few years.

Optical Music Recognition (OMR) is a field devoted to providing computers the ability to extract the musical content of a score from the optical scanning of its source. The output of an OMR system is the music score encoded in some structured digital format such as MusicXML, MIDI or MEI. Typically, the transcription of early music documents is treated differently with respect to conventional OMR methods due to specific features (for instance, the different notation or the quality of the document). Although there exist several works focused on early music documents transcription [9,10], the specificity of each type of notation or writing makes it difficult to generalize these developments. This is especially detrimental to the evolution of the field because it is necessary to implement new processing techniques for each type of archive. Even worse, new labelled data are also needed to develop techniques for automatic recognition, which might imply a significant cost.

Notwithstanding the efforts devoted to improving these systems, their performance is far from being optimal [12]. In fact, assuming that a totally accurate automatic transcription is not possible, and might never be, user-centred recognition is becoming an emergent framework. Instead of a fully-automatized process, *computer-aided* systems are being considered, with which the user collaborates actively to complete the recognition task [16].

The goal of this kind of systems is to facilitate the task for the user, since it is considered the most valuable resource [2]. In the case of the transcription of early music documents, the potential user is the expert musicologist who understands the meaning of any nuance that appears in the score. However, very often these users find the use of a pen more natural and comfortable than keyboard entry or drag-and-drop actions with the mouse. Using a tablet device and e-pen, it is possible to develop an ergonomic interface to receive feedback from users' drawings. This is specially true for score post-edition where the user, instead of sequentially inputting symbols has to correct some of them, and for that, direct manipulation is the preferred interaction style.

Such an interface could be used to amend errors made by the system in a simpler way for the user, as has been proposed for automatic text recognition [1]. However, there are studies showing that, when the task is too complex, users prefer to complete the task by themselves because the human-machine interaction is not friendly enough [14]. Therefore, this interface could also be used to develop a manual transcription system that would be more convenient and intuitive than conventional score editors. Moreover, this transcription system might be useful in early stages of an OMR development, as it could be used to acquire training data more efficiently and ergonomically, which is specially interesting for old music notations.

Unfortunately, although the user is provided with a more friendly interface to interact with the system, the feedback input is not deterministic this way. 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.

For all the reasons above, this article presents our research on the capabilities of musical notation recognition with a system whose input is a pen-based interface. To this end, we shall assume a framework in which the user traces symbols on the score, regardless of the purpose of this interaction (OMR error correction, digitizing the content, acquire labelled data, etc.). As a result, the system 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* mode. One of the main hypothesis of this study is that the combination of both modalities leads to better results than using just either the pen data or the symbol image.

The rest of the paper is structured as follows: Section 2 introduces the corpora collected and utilized, which comprises data of Spanish early music written in White Mensural notation; Section 3 describes a multimodal classifier that exploits both offline and online data; Section 4 presents the results obtained with such classifier; and Section 5 concludes the present work.

## 2. MULTIMODAL DATA COLLECTION

This work is a first seed of a case study to digitize a historical musical archive of early Spanish music. The final objective of the whole project is to encode the musical content of a huge archive of manuscripts dated between 16th and 18th centuries, handwritten in mensural notation, in the variant of the Spanish notation at that time [5]. A short sample of a piece from this kind of document is illustrated in Figure 1.

This section describes the process developed to collect multimodal data of isolated musical symbol from images of scores. A massive collection of data will allow us to develop a more effective classification system and to go deeper into the analysis of this kind of interaction. Let us note that the important point in our interactive system is to better understand user actions. While a machine is assumed to make some mistakes, it is unacceptable to force the user to draw the same symbol of score many times. To this end, our intention is to exploit both offline data (image) and online data (e-pen user tracing) received.

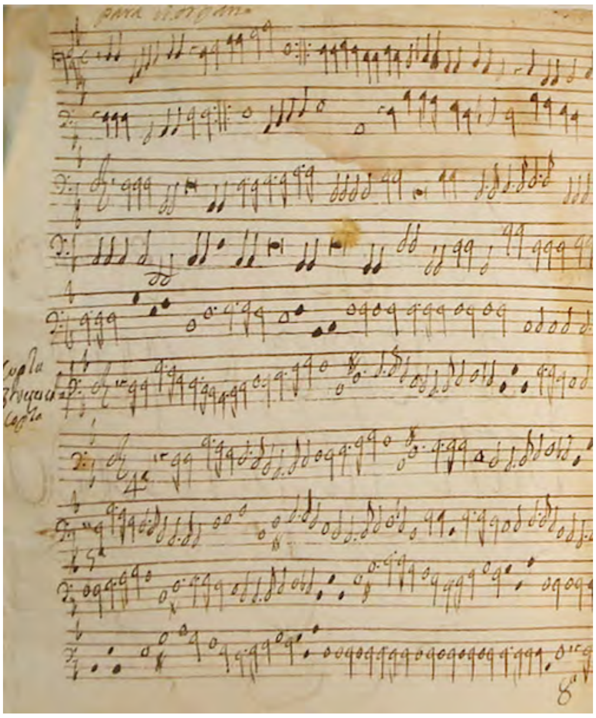
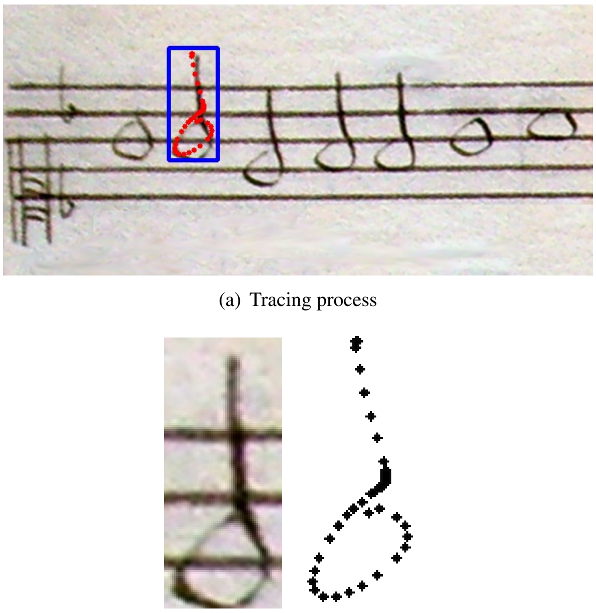


**Figure 1**. Example of page of a music book written in handwritten white mensural notation from Spanish manuscripts of centuries 16th to 18th.

Our idea is to simulate the same scenario of a real application. Therefore, we loaded the images of the scores on a digital surface to make users trace the symbols using the electronic pen. The natural symbol isolation of this kind of input is the set of strokes —data collected between pen-down and pen-up actions. To allow tracing symbols with several strokes, a fixed elapsed time is used to detect when a symbol has been completed. If a new stroke starts before this time lapse, it is considered to belong to the same symbol than the previous one.

Once online data is collected and manually grouped into symbol classes, the offline data is also extracted from this information. A bounding box is obtained from each group of strokes belonging to the same symbol, storing the maximum and minimum values of each coordinate (plus a small margin) among all the trace points collected. This bounding box indicates where the traced symbol can be found in the image. Therefore, with the sole effort of the tracing process, both online and offline data are collected. Note that the extraction of the offline data is driven by the tracing process, instead of deciding at every moment the bounds of each symbol.

Figure 2 illustrates the process explained above for a single symbol. Although the online data is drawn in this example, the actual information stored is the sequence of 2D points in the same order they were collected, indicating



(a) Tracing process

(b) Offline data (c) Online data

**Figure 2**. 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 from the same sample.

the path followed by the e-pen.

Following this approach, several advantages are found: the final effort of collecting multimodal data is halved, since the online data collection simultaneously provides the offline data collection; the collected data mimics the scenario that might be found in the final application, when the user interacts with the machine; and the process becomes more user-friendly, which usually leads to a lower number of errors.

The collection was extracted by five different users from 70 different musical scores of different styles from the Spanish white mensural notation of 16th-18th centuries. The *Samsung Galaxy Note Pro 12.2* device (247 ppi resolution) was used and symbols were written by means of the stylus *S-Pen*. All the score images used are in the same scale, in which staff lines spacing is about 24 DP. [1] Due to the irregular conditions of the documents, this value is approximate but it can be used for normalizing with respect to other scores.

The obtained dataset consists of 10230 samples, each of which contains both a piece of image and the strokes followed during its tracing. These samples are spread over 30 classes. Table 1 lists the set of labels, including a typographic example and the number of samples per each. The number of symbols of each class is not balanced but it depicts the same distribution found in the documents.

Every symbol that must be differentiated for preservation purposes was considered as a different class. For

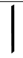
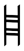
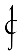
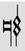
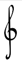
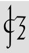
| Label | Image | Count |
|---|---|---|
| barline |  | 46 |
| brevis | | 210 |
| coloured brevis | | 28 |
| brevis rest | | 171 |
| c-clef |  | 169 |
| common time | | 29 |
| cut time |  | 56 |
| dot | | 817 |
| double barline | | 73 |
| custos | | 285 |
| f-clef 1 | | 52 |
| f-clef 2 |  | 43 |
| fermata | | 75 |
| flat | | 274 |
| g-clef |  | 174 |
| beam | | 85 |
| longa | | 30 |
| longa rest | | 211 |
| minima | | 2695 |
| coloured minima | | 1578 |
| minima rest | | 427 |
| proportio minor | | 28 |
| semibrevis | | 1109 |
| coloured semibrevis | | 262 |
| semibrevis rest | | 246 |
| semiminima | | 328 |
| coloured semiminima | | 403 |
| semiminima rest | | 131 |
| sharp | | 170 |
| proportio maior |  | 25 |

**Table 1**. Details of the dataset obtained through the tracing process over 70 scores (images from 'Capitán' font).

[1] DP stands for *device independent pixels* in (Android) mobile application development

instance, there are two *f-clef* types because the graphical symbol is quite different despite having the same musical meaning. However, the orientation of the symbols does not make a different class since the same graphical representation with a vertical inversion can be found. In the case it was needed, the orientation could be obtained through an easy post-processing step.

We are making the dataset freely available at `http://grfia.dlsi.ua.es/`, where more information about the acquisition and representation of the data is detailed.

## 3. MULTIMODAL CLASSIFICATION

This section provides a classification experiment over the data described previously. Two independent classifiers are proposed that exploit each of the modalities presented by the data. Eventually, a late-fusion classifier that combines the two previous ones will be considered.

Taking into account the features of our case of study, an instance-based classifier was considered. Specifically, the Nearest Neighbour (NN) rule was used, as it is one of the most common and effective algorithms of this kind [3]. The choice is justified by the fact that it is specially suitable for interactive scenarios like the one found in our task: it is naturally adaptive, as the simple addition of new prototypes to the training set is sufficient (no retraining is needed) for incremental learning from user feedback. The size of the dataset can be controlled by distance-based data reduction algorithms [7] and its computation time can be improved by using fast similarity search techniques [17].

Decisions given by NN classifiers can be mapped onto probabilities, which are needed for the late fusion classifiers. Let $\mathcal{X}$ be the input space, in which a pairwise distance $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is defined. Let $\mathcal{Y}$ be the set of labels considered in the classification task. Finally, let $T$ denote the training set of labelled samples $\{(x_i, y_i) : x_i \in \mathcal{X}, y_i \in \mathcal{Y}\}_{i=1}^{|T|}$.

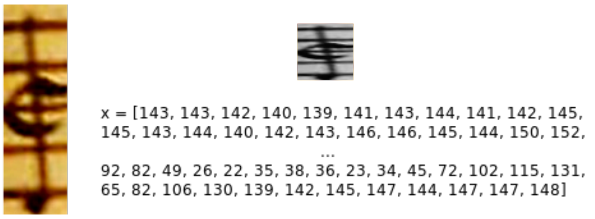
Let us now assume that we want to know the posterior probability of each class $y \in \mathcal{Y}$ for the input point $x \in \mathcal{X}$ ($P(y|x)$) following the NN rule. A common estimation makes use of the following equations [4]:

$$p(y|x) = \frac{1}{\min_{(x',y') \in T : y'=y} d(x, x') + \epsilon} \quad (1)$$

$$P(y|x) = \frac{p(y|x)}{\sum_{y' \in \mathcal{Y}} p(y'|x)}, \quad (2)$$

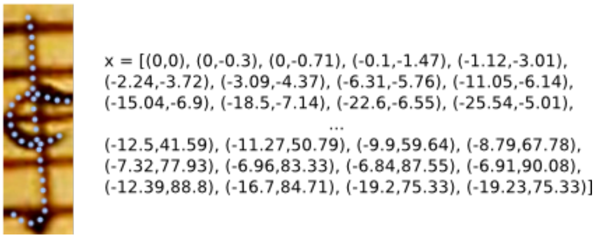
where $\epsilon$ is a negligible value used to avoid infinity calculations. That is, the probability of each class is defined as the inverse of the distance to the nearest sample of that class in the training set. Note that the second term is used to ensure that the sum over the probability of each class is 1. Finally, the decision $\hat{y}$ of the classifier for an input $x$ is given by a *maximum a posteriori* criterion:

$$\hat{y} = \arg\max_{y} P(y|x) \quad (3)$$



```
x = [143, 143, 142, 140, 139, 141, 143, 144, 141, 142, 145,
145, 143, 144, 140, 142, 143, 146, 146, 145, 144, 150, 152,
...
92, 82, 49, 26, 22, 35, 38, 36, 23, 34, 45, 72, 102, 115, 131,
65, 82, 106, 130, 139, 142, 145, 147, 144, 147, 147, 148]
```

**Figure 3**. Offline modality of a *cut time* symbol for classification: feature vector containing the greyscale value of each position of the rescaled image.



```
x = [(0,0), (0,-0.3), (0,-0.71), (-0.1,-1.47), (-1.12,-3.01),
(-2.24,-3.72), (-3.09,-4.37), (-6.31,-5.76), (-11.05,-6.14),
(-15.04,-6.9), (-18.5,-7.14), (-22.6,-6.55), (-25.54,-5.01),
...
(-12.5,41.59), (-11.27,50.79), (-9.9,59.64), (-8.79,67.78),
(-7.32,77.93), (-6.96,83.33), (-6.84,87.55), (-6.91,90.08),
(-12.39,88.8), (-16.7,84.71), (-19.2,75.33), (-19.23,75.33)]
```

**Figure 4**. Online modality of a *cut time* symbol for classification: sequence of coordinates indicating the path followed by the e-pen during the tracing process.

### 3.1 Offline classifier

The offline classifier takes the image of a symbol as input. To simplify the data, the images are converted to greyscale. Then, since they can be of different sizes, a fixed resizing process is performed, in the same way that can be found in other works, like that of Rebelo et al. [11]. At the end, each image is represented by an integer-valued feature vector of equal length that stores the greyscale value of each pixel (see Figure 3). Over this data, Euclidean distance can be used for the NN classifier. A preliminary experimentation fixed the size of the images to 30 × 30 (900 features), although the values within the configurations considered did not vary considerably.

### 3.2 Online classifier

In the online modality, the input is a series of 2D points that indicates the path followed by the pen (see Figure 4). It takes advantage of the local information, expecting that a particular symbol follows similar paths. The information contained in this modality provides a new perspective on the recognition and it does not overlap with the nature of the offline recognition.

The digital surface collects the strokes at a fixed sampling rate so that each one may contain a variable number of points. However, several distance functions can be applied to this kind of data. Those considered in this work are the following:

- Dynamic Time Warping (DTW) [15]: a technique for measuring the dissimilarity between two time signals which may be of different duration.
- Edit Distance with Freeman Chain Code (FCC): the sequence of points representing a stroke is converted

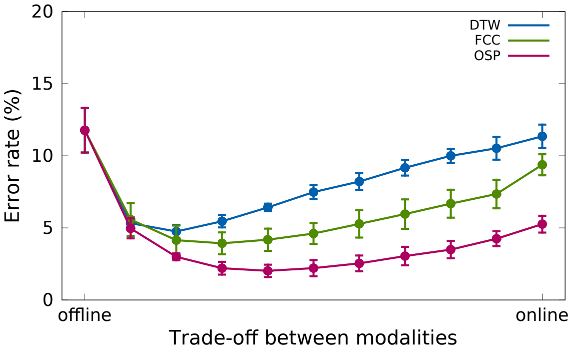


**Figure 5**. Average results with respect to the weight ($\alpha$) given to each modality for the configurations considered, from *offline* ($\alpha = 0$) to *online* ($\alpha = 1$).

into a string using a codification based on Freeman Chain Code [6]. Then, a Edit Distance [8] can be applied to measure distance.

- Edit Distance for Ordered Set of Points (OSP) [13]: an extension of the Edit Distance for its use over ordered sequences of points, such those collected by the e-pen.

### 3.3 Late-fusion classifier

A straightforward late fusion has been used here. The idea is to combine linearly the decisions taken by the two base classifiers. That is, probabilities of individual classifiers are combined by a weighted average:

$$P_{\text{fusion}}(y|x) = \alpha \cdot P_{\text{on}}(y|x) + (1-\alpha) \cdot P_{\text{off}}(y|x) \quad (4)$$

where $P_{\text{off}}$ and $P_{\text{on}}$ denote the probabilities obtained by offline and online classifiers, respectively. A parameter $\alpha \in [0,1]$ is established to tune the relevance given to each modality. We will consider several values of $\alpha$ ranging from 0 to 1 during experimentation.

## 4. EXPERIMENTATION

Experimentation followed a 10-fold cross-validation scheme. The independent folds were randomly created with the sole constraint of having the same number of samples per class (where possible) in each of them. All the dissimilarities described in Section 3 for the online classifier will be tested.

Table 2 illustrates the error rate (%) achieved with respect to $\alpha$ for this experiment. Note that $\alpha = 0$ column yields the results of the offline classifier as well as $\alpha = 1$ is equal to the online classifier. A summary of the average results is also illustrated in Figure 5.

An initial remark to begin with is that the worst results of the late-fusion classifiers are achieved when each is modality is used separately, with an average error of 11.77 for the offline modality and of 11.35, 9.38 and 5.26 for DTW, FCC and OSP, respectively. Not surprisingly, best results are those that combine both natures of the data, satisfying the hypothesis that two signals are better than one.

| $\alpha$ | DTW | FCC | OSP |
|---|---|---|---|
| 0.0 | 11.8 ± 1.5 | 11.8 ± 1.5 | 11.8 ± 1.5 |
| 0.1 | 5.3 ± 0.4 | 5.5 ± 1.1 | 4.9 ± 0.7 |
| 0.2 | **4.7** ± **0.4** | 4.1 ± 1.0 | 3.0 ± 0.2 |
| 0.3 | 5.4 ± 0.4 | **3.9** ± **0.8** | 2.2 ± 0.4 |
| 0.4 | 6.4 ± 0.3 | 4.1 ± 0.7 | **2.0** ± **0.4** |
| 0.5 | 7.4 ± 0.5 | 4.6 ± 0.7 | 2.2 ± 0.5 |
| 0.6 | 8.2 ± 0.6 | 5.2 ± 0.9 | 2.5 ± 0.5 |
| 0.7 | 9.1 ± 0.5 | 5.9 ± 1.0 | 3.0 ± 0.6 |
| 0.8 | 9.8 ± 0.5 | 6.6 ± 0.9 | 3.4 ± 0.6 |
| 0.9 | 10.5 ± 0.8 | 7.3 ± 0.9 | 4.2 ± 0.5 |
| 1.0 | 11.3 ± 0.8 | 9.3 ± 0.7 | 5.2 ± 0.5 |

**Table 2**. Error rate (average ± std. deviation) obtained for a 10-fold cross validation experiment with respect to the value used for tuning the weight given to each modality ($\alpha$) and the distances for the online modality (DTW, FCC and OSP). Bold values represent the best average result for each configuration considered.

Results also report that the tuning of $\alpha$ is indeed relevant since it makes the error vary noticeably. An interesting point to mention is that, although the online modality is more accurate than the offline one by itself, the best tuning in each configuration always gives more importance to the latter. This might be caused by the lower variability in the writing style of the original scribes.

The best results, on average, are reported by the late-fusion classifier considering OSP distance for the online modality, with an $\alpha = 0.4$. In such case, just 2 % of error rate is obtained, which means that the interaction is well understood by the system in most of the cases. Note that a more comprehensive search of the best $\alpha$ may lead to a better performance —for instance, in the range $(0.3, 0.5)$— but the improvement is not expected to be significant.

Although the results report a fair accuracy, the use of semantic music models is expected to avoid some of these mistakes by using contextual information. Therefore, a nearly optimal performance could be obtained during the interaction with the user.

## 5. CONCLUSIONS

This paper presents a new approach to interact with musical documents, based on the use of an electronic pen. Our framework assumes that the user traces each musical symbol of the score, and the system receives a *multimodal* input accordingly: the sequence of coordinates indicating the trajectory of the e-pen (online mode) and the underlying image of the score itself (offline mode).

An interface based on this idea could be used in a number of contexts related to interact with music scores in a more intuitive way for the user. For instance, to amend OMR errors, to acquire training data in the early stages of the development, or even as a part of a complete manual transcription system.

This framework has been applied to a music archive of Spanish music from the 16th to 18th centuries, handwritten in white mensural, with the objective of obtaining data for our experiments. The result of processing this collection has been described and made available for research purposes.

Experimentation with this dataset is presented, considering several classifiers. The overall analysis of this experiments is 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 separately.

As a future line of work, the reported analysis will be used to build a whole *computer-aided* system, in which the user interacts with the system by means of an electronic pen to digitize music content. Since the late-fusion classifier is close to its optimal performance, it seems to be more interesting to consider the development of semantic models that can amend misclassifications by using contextual information (*e.g.*, a score starts with a clef). In addition, further effort is to be devoted to visualization and user interfaces.

## 6. ACKNOWLEDGEMENTS

This work has been supported by the Spanish Ministerio de Educación, Cultura y Deporte through a FPU fellowship (Ref. AP2012–0939) and the Spanish Ministerio de Economía y Competitividad through project TIMuL (No. TIN2013-48152-C2-1-R, supported by UE FEDER funds).

Authors would like to thank Ichiro Fujinaga for inspiring the title of this paper, and Antonio Ezquerro and Luis Antonio González (Instituci Milà i Fontanals, Spanish National Research Council), who allowed the utilization of the music sources for research purposes.

## 7. REFERENCES

[1] Vicent Alabau, Carlos D. Martínez-Hinarejos, Verónica Romero, and Antonio L. Lagarda. An iterative multimodal framework for the transcription of handwritten historical documents. *Pattern Recognition Letters*, 35:195–203, 2014.

[2] Jesús Andrés-Ferrer, Verónica Romero, and Alberto Sanchis. *Multimodal Interactive Pattern Recognition and Applications*, chapter General Framework. Springer, 1st edition edition, 2011.

[3] T. Cover and P. Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27, January 1967.

[4] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. John Wiley & Sons, New York, NY, 2nd edition, 2001.

[5] Antonio Ezquerro Esteban, editor. *Música de la Catedral de Barcelona a la Biblioteca de Catalunya*. Biblioteca de Catalunya, Barcelona, 2001.

[6] Herbert Freeman. On the encoding of arbitrary geometric configurations. *IRE Transactions on Electronic Computers*, EC-10(2):260–268, 1961.

[7] Salvador García, Julián Luengo, and Francisco Herrera. *Data Preprocessing in Data Mining*, volume 72 of *Intelligent Systems Reference Library*. Springer, 2015.

[8] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. Technical Report 8, 1966.

[9] João Rogério Caldas Pinto, Pedro Vieira, and João Miguel da Costa Sousa. A new graph-like classification method applied to ancient handwritten musical symbols. *IJDAR*, 6(1):10–22, 2003.

[10] Laurent Pugin. Optical music recognition of early typographic prints using hidden markov models. In *ISMIR 2006, 7th International Conference on Music Information Retrieval, Victoria, Canada, 8-12 October 2006, Proceedings*, pages 53–56, 2006.

[11] A. Rebelo, G. Capela, and Jaime S. Cardoso. Optical recognition of music symbols. *International Journal on Document Analysis and Recognition (IJDAR)*, 13(1):19–31, 2010.

[12] 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.

[13] Juan R. Rico-Juan and Jose M. Iñesta. Edit distance for ordered vector sets: A case of study. In *Structural, Syntactic, and Statistical Pattern Recognition*, volume 4109 of *Lecture Notes in Computer Science*, pages 200–207. Springer Berlin Heidelberg, 2006.

[14] V. Romero and J. Andreu Sanchez. Human Evaluation of the Transcription Process of a Marriage License Book. In *12th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1255–1259, Aug 2013.

[15] Hiroaki Sakoe and Seibi Chiba. Readings in Speech Recognition. In Alex Waibel and Kai-Fu Lee, editors, *Readings in Speech Recognition*, chapter Dynamic Programming Algorithm Optimization for Spoken Word Recognition, pages 159–165. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990.

[16] Alejandro Hector Toselli, Verónica Romero, Moisés Pastor, and Enrique Vidal. Multimodal interactive transcription of text images. *Pattern Recognition*, 43(5):1814–1825, 2010.

[17] Pavel Zezula, Giuseppe Amato, Vlastislav Dohnal, and Michal Batko. *Similarity Search - The Metric Space Approach*, volume 32 of *Advances in Database Systems*. Kluwer, 2006.