applied sciences

MDPI

Article

Exploiting the Two-Dimensional Nature of Agnostic Music Notation for Neural Optical Music Recognition

María Alfaro-Contreras *,† **and Jose J. Valero-Mas** †

Department of Software and Computing Systems, University of Alicante, Ctra. San Vicente del Raspeig s/n, 03690 Alicante, Spain; jjvalero@dlsi.ua.es

* Correspondence: malfaro@dlsi.ua.es; Tel.: +34-965-90-34-00

† The authors contributed equally to this work.

Abstract: State-of-the-art Optical Music Recognition (OMR) techniques follow an end-to-end or holistic approach, i.e., a sole stage for completely processing a single-staff section image and for retrieving the symbols that appear therein. Such recognition systems are characterized by not requiring an exact alignment between each staff and their corresponding labels, hence facilitating the creation and retrieval of labeled corpora. Most commonly, these approaches consider an agnostic music representation, which characterizes music symbols by their shape and height (vertical position in the staff). However, this double nature is ignored since, in the learning process, these two features are treated as a single symbol. This work aims to exploit this trademark that differentiates music notation from other similar domains, such as text, by introducing a novel end-to-end approach to solve the OMR task at a staff-line level. We consider two Convolutional Recurrent Neural Network (CRNN) schemes trained to simultaneously extract the shape and height information and to propose different policies for eventually merging them at the actual neural level. The results obtained for two corpora of monophonic early music manuscripts prove that our proposal significantly decreases the recognition error in figures ranging between 14.4% and 25.6% in the best-case scenarios when compared to the baseline considered.

Keywords: optical music recognition; deep learning; connectionist temporal classification; agnostic music notation; sequence labeling

Citation: Alfaro-Contreras, M.; Valero-Mas, J.J. Exploiting the Two-Dimensional Nature of Agnostic Music Notation for Neural Optical Music Recognition. *Appl. Sci.* **2021**, *11*, 3621. https://doi.org/10.3390/app11083621

Academic Editor: Junhong Park

Received: 22 February 2021
Accepted: 15 April 2021
Published: 17 April 2021

1. Introduction

Music is one of the cornerstones of cultural heritage [1]. Throughout history, the main means of transmitting and preserving this art has been its engravement in so-called music scores, i.e., documents in which music composers graphically encode a piece of music as well as the way to perform it [2]. This particular visual language is identified as music notation and is known to be considerably different depending on epoch, culture, instrumentation, and genre, among many others [3].

This engraving process was carried out manually until the end of the 20th century, mostly done by handwriting or typesetting processes; thus, millions of music documents are only available as physical documents [4]. While large quantities of them have been scanned into digital images, only a few hundreds have been stored in a structured digital format that allows tasks such as indexing, editing, or critical publication [5]. Since its manual transcription constitutes a costly, tedious, and prone-to-error task, the development of systems capable of automatically performing this process is of remarkable importance, even more so considering the vast amount of existing music archives [6]. Optical Music Recognition (OMR) is the research field that investigates how to computationally read music notation in printed documents and to store them in a digital structured format [7].

One of the main challenges in this field lies in the aforementioned heterogeneity of the documents, especially when considering the different notation styles present in historical documents [8]. Thus, while OMR has been an active research area for decades [9,10],

such particularity prevents the field from providing general developments for any type of document.

Traditionally, most OMR systems rely on a multi-stage pipeline, which, in a broad sense, is devoted to segmenting the complete score into single units (e.g., staves or symbols), which are eventually recognized [11] using traditional machine learning or signal processing techniques. However, it must be noted that such approaches achieve competitive recognition rates at the expense of using certain heuristics adequately designed for a particular notation and writing/printing style, being only applicable to the case for which they were devised [12]. In this regard, this framework becomes considerably limited in terms of scalability given that a new pipeline and a set of heuristics must be designed for each particular document and/or notation.

Recent developments in the so-called *deep learning* paradigm have led to a considerable renewal of learning-based approaches [13]. This area represents the current state-of-the-art in a wide range of applications, with *image classification* being a particular example in which traditional machine learning techniques seemed to have reached a glass ceiling [14]. Hence, it is not strange that OMR has severely benefited from that field to somehow palliate its issues.

One of the key advantages of deep learning is that it allows for the creation of the so-called *holistic* or *end-to-end* neural solutions to tackle recognition tasks, i.e., systems in which both feature extraction and classification processes are learned jointly [15]. This particularity prevents the need for designing an entire pipeline for each particular case of study since the traditionally hand-crafted tasks are directly inferred from the data itself. Such a functionality has been largely explored in the literature, with OMR being a case in which it has clearly proven its usefulness [16], at least in symbol recognition at the isolated staff level. It must be noted that, while generative methods based on Hidden Markov Models have also been considered for holistic OMR as in the works by Pugin [17] or Calvo-Zaragoza et al. [18], the performance of the commented neural methods outperforms them. Nevertheless, there is still room for improvement in neural-based holistic OMR by exploiting the inherent particularities of this type of data, which have not been totally explored.

When considering an *agnostic music representation* [19], i.e., a representation based on the graphical content rather than its musical meaning, symbols depict a two-dimensional nature. As a consequence, every single element reports two geometrical pieces of information at once [20]: the *shape*, which encodes the temporal duration of the event (sound or absence of it), and the *height*, which indicates the pitch of the event represented with its vertical position in the staff. However, when holistic OMR systems are trained, each possible combination of shape and height is represented as unique categories, which leads to music symbols being treated the same way as text characters [21]. Based on this premise, some recent research OMR works [22–25] have explored the aforementioned particularities of music notation, concluding that the individual exploitation of each dimension generally yields better recognition rates.

In this work, we propose to further explore and exploit the two-dimensional nature of music symbols and, more precisely, we focus on symbol recognition at a staff-line level of monophonic early music documents. Considering the end-to-end neural-based framework by Calvo-Zaragoza et al. [20] as a starting point, we propose and discuss the separate extraction of shape and height features and then propose several integration policies. The neural end-to-end approach presented in this work fills a gap in the related literature as this double dimension has not been previously addressed in this particular context, thus constituting a clear innovation in the OMR field. The results obtained for two different corpora show that our proposal outperforms the recognition performance of the baseline considered, even in cases in which there is considerably narrow room for improvement.

The rest of the paper is organized as follows: Section 2 overviews some related proposals for this topic; Section 3 thoroughly develops our proposed approach; Section 4 describes the experimental setup; Section 5 presents and analyzes the results; Section 6

provides a general debate of the insights obtained in the work; and finally, Section 7 concludes the present work, along with some ideas for future research.

2. Background

Even though the term OMR encompasses a vast range of scenarios—research might vary depending on the music notation system or the engraving mechanism of the manuscripts—most of the existing literature is framed within a multi-stage pipeline that divides the challenge into a series of independent phases [26]. Bainbridge and Bell [9] properly described and formalized the de facto standard workflow, which was later thoroughly reviewed by Rebelo et al. [10]. This sequential pipeline comprises four main blocks: (i) *image preprocessing*, which aims at palliating problems mostly related to the scanning process and paper quality; (ii) *symbol segmentation and classification*, which focuses on the detection and actual labeling of different elements of the image meant to be recognized; (iii) *reconstruction of the music notation*, which postprocesses the recognition process; and (iv) an *output encoding* stage that stores the recognized elements into a suitable symbolic format. In this work, we focus on the symbol segmentation and classification one.

The inclusion of deep learning strategies in the OMR field produced a shift towards the use of end-to-end or holistic systems for symbol recognition [27]. Some examples of work addressing the recognition process at a staff level may be found in the literature for common Western notation [12,16,28,29] as well as mensural [20], neumatic [30], and ancient organ tablature [31] handwritten music. While our proposal focuses on staff-line symbol recognition, it must be noted that research efforts are also devoted to addressing the issue of full-page recognition such as the proposal by Castellanos et al. [21].

In this context of staff-level holistic approaches for symbol recognition, most approaches rely on the use of architectures based on *Convolutional Recurrent Neural Networks* (CRNN). Such schemes work on the premise of using the convolutional stage to learn the adequate features for the case at issue, with the recurrent part being devoted to modeling the temporal (or spatial) dependencies of the symbols. This type of design has been mainly exploited using *Sequence-to-Sequence* (seq2seq) architectures [22,25] or considering the *Connectionist Temporal Classification* (CTC) training mechanism [12,16,20,28]. However, the work by Ríos-Vila et al. [25] shows that, when targeting the same handwritten corpus, CRNN-seq2seq models are not competitive against CRNN trained with CTC.

As previously introduced, music notation depicts a two-dimensional nature since each symbol is actually defined by a combination of a certain shape or glyph and its height or vertical position within the staff lines. While proven to be beneficial, this particularity has been rarely exploited in the literature, with some examples being the work by Nuñez-Alcover et al. [23], which shows that separately performing shape and height classification is beneficial in the context of isolated symbol classification of early music manuscripts; by van der Wel and Ullrich [22] and Ríos-Vila et al. [25], in which this conclusion was further reinforced in a context of CRNN-seq2seq models; and by Villarreal and Sánchez [24], which also proved the validity of such an exploitation by integrating the shape and height information at a Language Model level, which was based on Hidden Markov Models rather than on the actual optical estimation.

For all of the above, this paper presents a novel proposal that takes advantage of the two-dimensional nature of music symbols in the context of neural end-to-end OMR systems at the staff level. The basic premise in this case is that, instead of relying on a single CRNN scheme for processing each staff, we can devote two CRNN schemes to concurrently exploit the shape and height information and to eventually merge them at the actual neural level. As will be shown, this separate exploitation of the music symbol information provides a remarkable improvement in terms of recognition performance when compared to the baseline in which this duality is not considered.

3. Methodology

This section describes the recognition framework of the work and the different proposals for properly exploiting the dual nature of agnostic music notation.

The proposed OMR recognition task works at the staff level; thus, we assume that a certain preprocess (e.g., [32,33]) already segmented the different staves in a music sheet. Our goal is, given an image of a single staff, to retrieve the series of symbols that appear therein, i.e., our recognition model is a *sequence labeling* task [34].

Formally, let $\mathcal{X}$ represent a space of music staves. Additionally, let Σ represent a symbol vocabulary and $\mathcal{Z} = \Sigma^*$ be the complete set of possible sequences that may be obtained from that vocabulary.

Since we are dealing with a supervised learning task, we assume the existence of a set $\mathcal{T} = \{(x_i, \mathbf{z}_i) : x_i \in \mathcal{X},\ \mathbf{z}_i \in \mathcal{Z}\}_{i=1}^{|\mathcal{T}|}$, which relates a given staff x_i to the sequence of symbols $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iN})$. Furthermore, we assume that there is an underlying function $g : \mathcal{X} \to \mathcal{Z}$, which is the one we aim to approximate as $\hat{g}(\cdot)$.

In this work, given its competitive performance reported in the literature, we consider the introduced Convolutional Recurrent Neural Network (CRNN) scheme together with the Connectionist Temporal Classification (CTC) training algorithm [35] for approximating function $\hat{g}(\cdot)$. Based on this premise, we derive different neural designs for performing the recognition task.

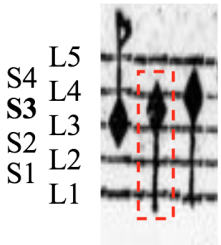
The rest of this section further develops the idea of the two-dimensional natural nature of music symbols when considering agnostic notation and its application in our case as well as the different neural architectures considered.

3.1. Symbol Representation

As commented upon, the agnostic notation considered allows for defining each music symbol using its individual shape and height graphical components. Note that this duality can be applied to all symbols, even to those that indicate the absence of sound, i.e., rests, because they may also appear at different vertical positions.

Let Σ_S and Σ_H be the spaces for the different shape and height labels, respectively. Formally, $\Sigma_T = \Sigma_S \times \Sigma_H$ represents the set of all possible music symbols in which a given ith element is denoted as the 2-tuple $\langle s_i, h_i \rangle : s_i \in \Sigma_S,\ h_i \in \Sigma_H$. However, while all aforementioned combinations are theoretically possible, in practice, some pairs are very unlikely to appear; hence, $\Sigma_T \subset \Sigma_S \times \Sigma_H$. In a practical sense, to facilitate the convergence of the model, we restrict the Σ_T vocabulary to the 2-tuples elements present in the corpus, i.e., $\Sigma_T = \{\Sigma_S, \Sigma_H\}^{\mathcal{T}}$.

Figure 1 shows a graphical example of the commented agnostic representation for a given symbol in terms of its shape (Σ_S), height (Σ_H), and combined labels (Σ_T).



Shape label: *note.quarter_down*
Height label: *S3*
Combined label: *note.quarter_down:S3*

Figure 1. Agnostic representation of a handwritten music symbol, showing its shape, height, and combined labels. Note that L*n* and S*n*, respectively, denote the line or space of the staff on which the symbol may be placed, which refers to its height property.

3.2. Recognition Architectures

The architecture of both recognition frameworks, baseline and proposed, are described below.

3.2.1. Baseline Approach

A CTC-trained CRNN is considered the state-of-the-art for the holistic approach in OMR, having been successfully applied in a number of work in the literature [12,16,20,28]. As aformentioned, this architecture models the posterior probability of generating a sequence of output symbols given an input image.

A CRNN is formed by an initial block of *convolutional* layers followed by another group of *recurrent* stages [28]. In such a particular configuration, the *convolutional* block is meant to learn adequate features for the case at issue, while the set of *recurrent* layers model the temporal or spatial dependencies of the elements from the initial feature-learning block.

The network is trained using the commented upon Connectionist Temporal Classification (CTC) training function [35], which allows for training the CRNN scheme using unsegmented sequential data. In our case, this means that, for a given staff image $x_i \in \mathcal{X}$, we only have its associated sequence of characters $\mathbf{z}_i \in \mathcal{Z}$ as its expected output, without any correspondence at the pixel level or similar input-output alignment. Due to its particular training procedure, CTC requires the inclusion of an additional "*blank*" symbol within the Σ vocabulary, i.e., $\Sigma' = \Sigma \cup \{blank\}$.

During the prediction or decoding phase, CTC assumes that the architecture contains a fully connected network with $|\Sigma'|$ outputs and a *softmax* activation. While there exist several possibilities for performing this inference phase [36], we resort to so-called *greedy* decoding for comparative purposes with the considered baseline works. Assuming that the recurrent layer outputs sequences of length K, this approach retrieves the most probable symbol per step. Equation (1) mathematically describes this process.

$$\boldsymbol{\pi} = \arg\max_{\boldsymbol{\pi} \in \Sigma'^K} \prod_{k=1}^{K} y^k_{\pi_k} \tag{1}$$

where $y^k_{\pi_k}$ represents the activation probability of symbol π_k and time-step k and where $\boldsymbol{\pi}$ is the retrieved sequence of length K.

Eventually, a $\mathcal{B}(\cdot)$ squash function that merges consecutive repeated symbols and removes the *blank* label is applied to the $\boldsymbol{\pi}$ sequence obtained as an output of the recurrent layer. Thus, the actual predicted sequence is obtained as $\mathbf{z}' = \mathcal{B}(\boldsymbol{\pi})$, where $|\mathbf{z}'| \leq K$.

In this work, we consider as the baseline the particular CRNN configuration proposed in the work by Calvo-Zaragoza et al. [20]. This neural model comprises four convolutional layers for the feature extraction process followed by two recurrent units for the dependency modeling. As commented upon, the output of the last recurrent layer is connected to a dense unit with $|\Sigma'|$ output neurons. A graphical representation of this configuration is depicted in Figure 2.

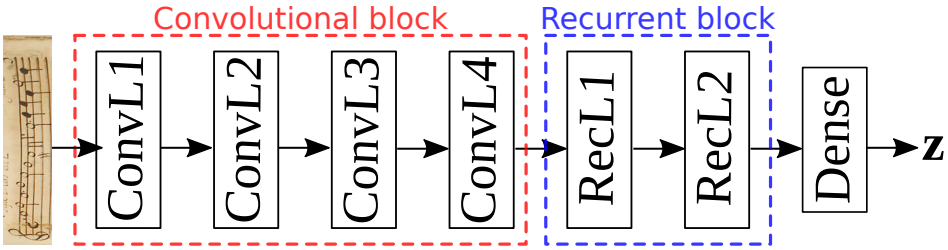


Figure 2. Graphical scheme of the CRNN configuration by Calvo-Zaragoza et al. [20].

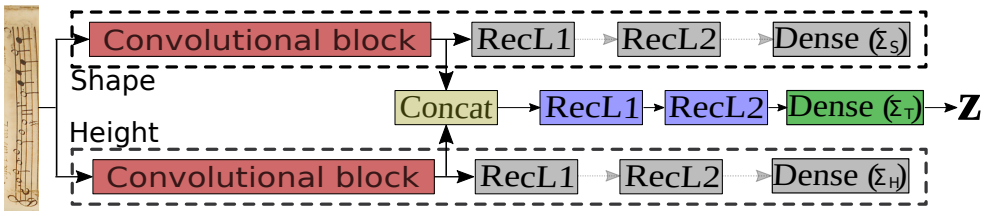
It must be finally noted that the different parameters for each layer is commented upon in Section 4, which deals with the actual experimentation carried out.

3.2.2. Proposed Approach

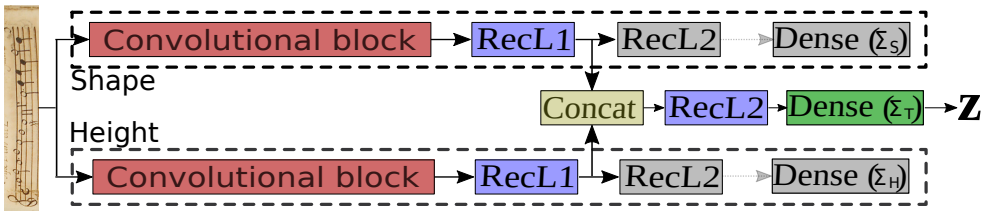
This section presents the different neural architectures proposed for exploiting the individual shape and height properties when considering an agnostic representation of music notation. For that, we modify the base CRNN architecture introduced in Section 3.2.1 by adding different layers to adequately exploit such pieces of information with the aim of improving the overall recognition rate.

More precisely, our hypothesis is that having two CRNN models that are specialized in retrieving the shape and height features may be beneficial with respect to having a unique system that deals with the task as a whole. The input staff image is individually processed by each model, and the different characteristics obtained by each single model may be gathered at some point of the neural model before the actual classification process.

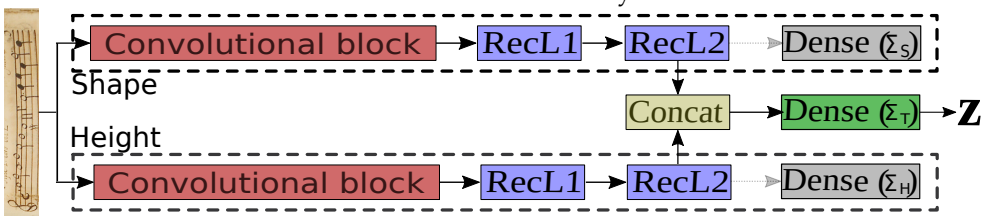
Based on that premise, we propose three different end-to-end architectures that basically differ on the point in which the two CRNN models are joined: (i) the *PreRNN* one, which joins the extracted features by each model right before the recurrent block; (ii) the *InterRNN* one, which performs this process after the first recurrent layer; and (iii) the *PostRNN* one, which gathers both sources of information after the recurrent block. These proposals are graphically shown in Figure 3.



(**a**) PreRNN model: features extracted by each specialized model are joined after the convolutional block.



(**b**) InterRNN model: features extracted by each specialized model are joined after the first recurrent layer.



(**c**) PostRNN model: features extracted by each specialized model are joined after the recurrent block.

Figure 3. Graphicaldescription of the three CRNN-based architectures proposed. The *Concat* block concatenates the input features from each branch. Gray layers represent the ones that are only considered during the training stage of the model.

As it may be noted, all models depict three differentiated parts. Out of these three parts, two of them constitute complete CRNN models specialized on a certain type of information (either shape or height) while the third one is meant to join the previous sources of information for eventual classification. These parts are thoroughly described below:

- Shape branch: a first CRNN model that focuses on the holistic classification of musical symbols in terms of their shape labels. Its input is the initial staff $x \in \mathcal{X}$, while its output is a sequence $\mathbf{z}^s \in \Sigma_S^*$ of shape symbols.

- Height branch: the other CRNN model devoted to recognition of the vertical position labels. In this case, a sequence $\mathbf{z}^h \in \Sigma_H^*$ of height symbols is retrieved out of the initial staff $x \in \mathcal{X}$.
- Combined branch: the one that combines the extracted features of the other two branches to perform joint estimation of music symbols in terms of their combined <shape:height> labels. Thus, given an initial input staff $x \in \mathcal{X}$, the branch retrieves a sequence $\mathbf{z} \in \Sigma_T^*$ of combined labels.

Note that all branches are separately trained using the same set of staves $\mathcal{T}$ with the CTC learning algorithm, simply differing on the output vocabulary considered. This way, we somehow bias the different *shape* and *height* CRNN branches to learn specific features for those pieces of information, whereas in the case of the *combined* branch, the training stage is expected to learn how to properly merge those separate pieces of information.

In a practical sense, in this work, we consider two different policies for training the introduced models: a first scenario in which the parameters of the entire architecture are learned from scratch, i.e., we sequentially train the different branches without any particular initialization, and a second case in which the shape and height branches are separately trained and, after their convergence, the same procedure of the first scenario is reproduced. In this regard, we may assess how influential the initial training stage of the different branches of the scheme is on the overall performance of the system.

It must be remarked that all proposed architectures perform the feature-merging operation after the convolutional block, having discarded other policies which may affect this stage. The reason for this is that, since the convolutional block is responsible for extracting the appropriate features out of the input image staff, we consider that the potential benefit may be achieved by merging those specialized characteristics in the recurrent stage.

While these new architectures suppose an increase in the network complexity with respect to the base model considered, separate exploitation of the graphic components of the commented agnostic music notation is expected to report an improvement in terms of recognition. Nevertheless, this increase does not imply a need for using more data since each part of the network specializes in a set of image features. Finally, note that, while the training stage may require more time until convergence, the inference phase spans practically the same time-lapse as in the base network since the different branches work concurrently before the merging phase.

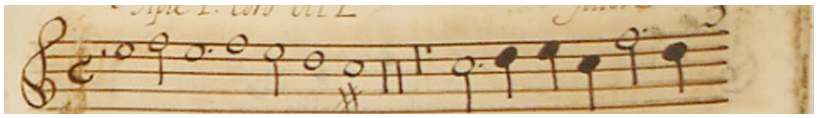
4. Experimental Setup

This section introduces the different corpora considered for assessing the goodness of our proposal as well as the actual definitions of the different layers of the neural models posed and the evaluation protocol contemplated.

4.1. Corpora

We considered two corpora of music scores depicting Mensural notation (the notation system used for the most part during the 16th and 17th centuries in Western music tradition) with varying print styles:

- *Capitan* corpus [37]: a manuscript of ninety-six pages dated from the 17th century of *missa* (sacred music). An example of a particular staff from this corpus is depicted in Figure 4a.
- *Il Lauro Secco* corpus [38]: a collection of one hundred and fifty-five typeset pages corresponding to an anthology of Italian madrigals of the 16th century. Figure 4b shows a staff example of this set.



(**a**) Handwritten music staff extracted from the *Capitan* set.

(**b**) Staff from the *Il Lauro Secco* corpus.

Figure 4. Music excerpts of the two corpora used in the experiments.

It must be mentioned that both corpora have already been considered in holistic staff-level symbol recognition. Thus, since our proposal addresses this same task, we segmented the initial music sheets as in the referenced works. A summary of the characteristics of these corpora is given in Table 1.

Table 1. Details of the corpora in terms of the number of staves, the average sequence length per staff, and the cardinality of the vocabulary for each label space considered.

Corpus	Staves	Avg. Staff Length	Vocabulary		
			Shape (Σ_S)	Height (Σ_H)	Combined (Σ_T)
Capitan	704	24.3	53	16	320
Il Lauro Secco	1100	28.7	33	14	182

Finally, for comparative purposes with the reference work, we reproduced the exact experimentation conditions. In this sense, we resized each image to a height of 64 pixels, maintaining the aspect ratio (thus, each sample might differ in width) and converted them to grayscale, with no further preprocessing. In terms of data partitioning, we also reproduced their train, validation, and test divisions with the same 5-fold cross validation policy.

4.2. Neural Network Configuration

As mentioned in Section 3.2, the different neural models proposed in this work for exploiting the commented upon duality of agnostic notation are based on the architecture of Calvo-Zaragoza et al. [20], which we considered our baseline. In this sense, while the configuration was broadly described in Section 3.2.1, the actual composition of each layer is depicted in Table 2.

Table 2. Layer-wise description for the CRNN architecture considered. Notation: $Conv(f, w \times h)$ stands for a convolution layer of f filters of size $w \times h$ pixels, *BatchNorm* performs the normalization of the batch, $LeakyReLU(\alpha)$ represents a leaky rectified linear unit activation with negative slope value of α, $MaxPool2D(w \times h, a \times b)$ stands for the max-pooling operator of dimensions $w \times h$ pixels with $a \times b$ striding factor, and $BLSTM(n, d)$ denotes a bidirectional long short-term memory unit with n neurons and d dropout value parameters.

Convolutional Block			
ConvL1	**ConvL2**	**ConvL3**	**ConvL4**
Conv2D (64, 5 × 5)	Conv2D (64, 5 × 5)	Conv2D (128, 3 × 3)	Conv2D (128, 3 × 3)
BatchNorm	BatchNorm	BatchNorm	BatchNorm
LeakyReLU (0.20)	LeakyReLU (0.20)	LeakyReLU (0.20)	LeakyReLU (0.20)
MaxPool2D (2 × 2, 2 × 2)	MaxPool2D (2 × 1, 2 × 1)	MaxPool2D (2 × 1, 2 × 1)	MaxPool2D (2 × 1, 2 × 1)
Recurrent Block			
	RecL1	**RecL2**	
	BLSTM (256, 0.5)	BLSTM (256, 0.5)	

In terms of the models proposed in Section 3.2.2, it must be noted that the nomenclature considered is the same as the one used for the baseline architecture. Hence, Table 2 also describes the different layers of the those proposed models.

Finally, these architectures were trained using the backpropagation method provided by CTC for 300 epochs using the ADAM optimizer [39] with a fixed learning rate. Batch size was fixed to 16 in the case of the *Capitan* corpus, while for the *Il Lauro Secco*, this value was set to 8.

4.3. Evaluation Protocol

We considered two figures of merit to measure the performance of the proposal. For their definition, let us assume a set $\mathcal{S} = \{(x_i, \mathbf{z}_i) : x_i \in \mathcal{X},\ \mathbf{z}_i \in \mathcal{Z}\}_{i=1}^{|\mathcal{S}|}$ of test data drawn from the same distribution as the train data $\mathcal{T}$ but disjoint from it. Considering that, the two metrics are described below:

- **Symbol Error Rate** (Sym-ER): computed as the average number of elementary editing operations (insertions, deletions, or substitutions) necessary to match the sequence predicted by the model with the ground truth sequence and normalized by the length of the latter. Mathematically, this is represented as follows:

$$\text{Sym-ER } (\%) = \frac{\sum_{i=1}^{|\mathcal{S}|} \text{ED}\Big(\mathbf{z}_i,\ \hat{h}(x_i)\Big)}{\sum_{i=1}^{|\mathcal{S}|} |\mathbf{z}_i|} \tag{2}$$

  where $\text{ED}(\cdot, \cdot)$ represents the string edit distance [40].

- **Sequence Error Rate** (Seq-ER): ratio of incorrectly predicted sequences, i.e., the percentage of sequences of recognized symbols that have at least one error. Mathematically, this may be represented as follows:

$$\text{Seq-ER } (\%) = \frac{\Big|\Big\{x \in \mathcal{S} : h(x) \neq \hat{h}(x)\Big\}\Big|}{|\mathcal{S}|} \tag{3}$$

It must be noted that, while these figures of merit may be applied to any predicted sequence disregarding the label space considered, in the particular case of the Σ_T^* space of combined shape and height classes, some additional insights may be obtained by further analyzing the results. As commented upon in Section 3.1, each symbol in the Σ_T vocabulary is a 2-tuple $\langle s_i, h_i \rangle : s_i \in \Sigma_S,\ h_i \in \Sigma_H$. Thus, for a given sequence $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iN}) \in \Sigma_T^*$, we may decouple each symbol into its individual shape $s_i \in \Sigma_S$ and height $h_i \in \Sigma_H$ components and assess the model in terms of the sole prediction of those individual elements while trained for the combined prediction.

5. Results

Having introduced the considered experimental scheme and the different models proposed, we now present and discuss the results obtained. Since the experiments have been performed in a cross-validation scheme, this section provides the average values obtained for each of the cases considered. It must be remarked that the figures reported constitute that of the test data partition for the case in which the validation data achieve their best performances.

The results obtained in terms of the Symbol Error Rate (Sym-ER) for the base neural configuration and the three different proposed architectures for each data corpus considered are shown in Figure 5. Note that these recognition models consider the Σ_T vocabulary case, i.e., each detected element is represented as the 2-tuple that defines both the shape and pitch of the symbol.

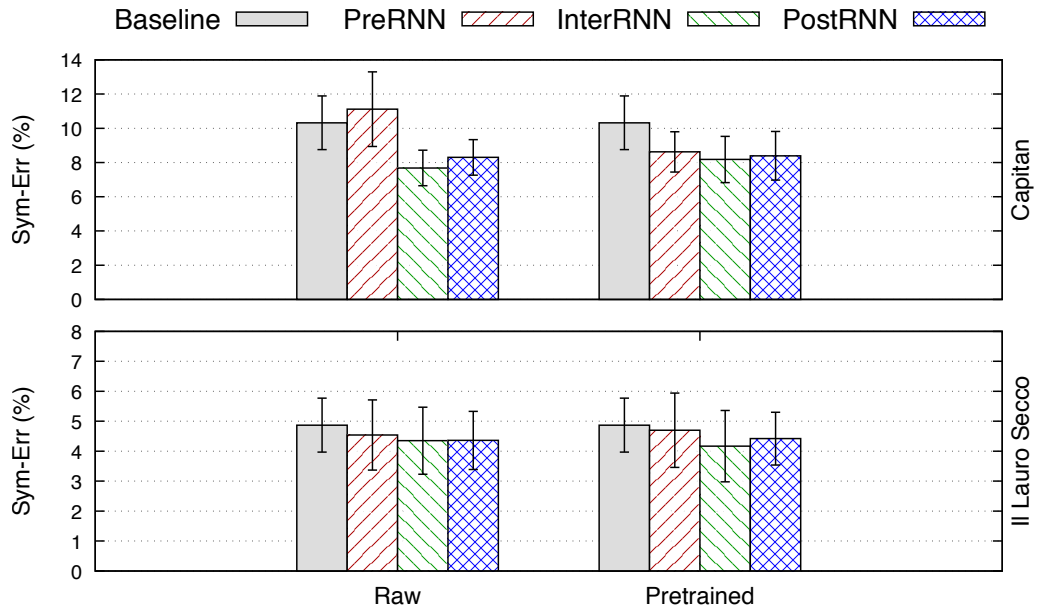


Figure 5. Comparison of the base neural model and the different architectures proposed for the Σ_T vocabulary case in terms of the Symbol Error Rate (Sym-ER (%)) and their respective standard deviations. *Pretrained* and *raw* denote the cases in which shape and height branches have been trained or not before considering the joint model, respectively.

An initial remark is that, as it can be appreciated in Figure 5, all proposed architectures improve the results obtained by their respective baselines except for the case of the *PreRNN* architecture in the *raw* scenario for the *Capitan* corpus. By examining this graph, we may also check that the *InterRNN* model is the one that consistently achieves the best error rates for all scenarios considered, tied only with the *PostRNN* proposal for the *raw* case of the *Il Lauro Secco* set.

It must be also noted that, for all cases, the error rates obtained with the *Capitan* corpus are remarkably higher than the ones obtained with the *Il Lauro Secco* set. This is a rather expected result since the graphical variability inherent to handwritten data compared to the typeset format supposes a drawback to the recognition algorithm. Moreover, the latter corpus depicts a lower vocabulary size, which also simplifies the complexity of the task.

Having visually checked the behavior of the different schemes considered, Table 3 numerically reports these Symbol Error Rate (Sym-ER) figures as well as the Sequence Error Rate (Seq-ER) ones for the same scenarios considered for a more detailed analysis.

Table 3. Results obtained in terms of the Symbol (Sym-ER) and Sequence (Seq-ER) Error Rates comparing the base neural model with the different architectures proposed for the Σ_T vocabulary case. *Pretrained* and *raw* denote the cases in which shape and height branches have been trained or not before considering the joint model, respectively.

	Baseline	**PreRNN**		**InterRNN**		**PostRNN**	
		Raw	**Pretrained**	**Raw**	**Pretrained**	**Raw**	**Pretrained**
Sym-ER (%)							
Capitan	10.32	11.12	8.62	7.68	8.18	8.30	8.39
Il Lauro Secco	4.87	4.54	4.70	4.35	4.17	4.36	4.42
Seq-ER (%)							
Capitan	86.53	89.37	81.39	77.15	76.41	78.21	80.70
Il Lauro Secco	36.16	33.40	37.47	31.26	32.75	31.45	34.34

The trends observed in the sequence error rate are consistent with the ones already provided by the symbol error rate metric in Figure 5. In general, for this Seq-ER figure of

merit, all results improve the base neural configuration considered except for the *PreRNN* model in two particular situations: when tackling the *Capitan* set in the *raw* scenario and for the *Il Lauro Secco* corpus in the *pretrained* case. The two other models, as commented upon, consistently outperform the baseline considered in all metrics and scenarios posed, with the *InterRNN* case being the one achieving slightly better results than the *PostRNN* one. A last point to comment upon from these figures is that, while differences between the *raw* and *pretrained* cases do not generally differ in a remarkable sense, convergence is always faster in the latter approach than in the former one.

We now further analyze this *InterRNN* model since, as commented upon, it is the one showing the best overall performance. To do so, we compare its performance when decoupling the predicted sequence of labels into their shape and height components against the base neural architecture when trained specifically for the Σ_S and Σ_H vocabularies. Table 4 provides the results of this analysis and reports the figures obtained for the baseline model.

Table 4. Detailed comparison in terms of the Symbol (Sym-ER) and Sequence (Seq-ER) Error Rates of the *InterRNN* proposal (denoted as *best model*) against the baseline case as well as the cases in which the base neural configuration is trained using the Σ_S and Σ_H vocabularies. Note that, for both the *baseline* and *best model* cases, we report the error rates when the predicted labels are decoupled in terms of their shape and height base components.

	Baseline			**Best Model**			**Shape**	**Height**
	$\boldsymbol{\Sigma_T}$	$\boldsymbol{\Sigma_S}$	$\boldsymbol{\Sigma_H}$	$\boldsymbol{\Sigma_T}$	$\boldsymbol{\Sigma_S}$	$\boldsymbol{\Sigma_H}$	$\boldsymbol{\Sigma_S}$	$\boldsymbol{\Sigma_H}$
Sym-ER (%)								
Capitan	10.32	7.09	8.84	7.68	4.57	6.59	4.92	7.52
Il Lauro Secco	4.87	4.05	4.10	4.17	3.37	3.48	2.14	2.47
Seq-ER (%)								
Capitan	86.53	77.92	83.75	77.15	57.05	72.16	61.57	78.67
Il Lauro Secco	36.16	28.85	29.83	32.75	25.61	26.53	24.20	25.13

As it may be checked, the results comparing the *baseline* and *best model* architectures when decoupling the combined labels into their shape and height components depict the expected behavior: for all cases and metrics, the *best model* decreases the error rate with respect to *baseline*. In the particular case of *Capitan*, the Symbol Error Rate decreases approximately by 2% while, for the *Il Lauro Secco*, this improvement is around 0.7% for both the shape and height individual vocabularies. The same trend is observed for the Sequence Error Rate, with the error decrease in the case of the shape vocabulary space for the *Capitan* corpus is especially noticeable, where the improvement is around 20%.

We now compare the performance of the *best model* configuration with that of the *shape* and *height* models. Regarding the *Capitan* set, the error rates achieved when decoupling the resulting predictions into the Σ_S and Σ_H vocabularies are lower than those of the actual dedicated models. These results suggest that the configuration exploits the individual sources of information in a synergistic manner that leads to such an improvement. Nonetheless, in the case of the *Il Lauro Secco* corpus, since the *shape* and *height* architectures achieve better recognition rates than those of the decoupled evaluation, there is still some room for improvement, which suggests that this may have not been the best architecture for this particular set.

Finally, let us comment that the narrow improvement margin that the different corpora depict must be taken into consideration. Focusing on the *Capitan* corpus, the baseline model depicts an initial Symbol Error Rate of 10.32%, which is reduced to 7.68% with the proposed architecture. While the absolute error reduction is 2.64%, it must be noted that it supposes an improvement of 25.6% with respect to the initial figure. Similarly, the *Il Lauro Secco* improves from an initial value of 4.87% to 4.17%, which is an absolute improvement of 0.7% but implies an error reduction of 14.4% with respect to that of the baseline model.

Statistical Significance Analysis

While the results obtained report that there is an improvement in the overall performance of the recognition task, at least when considering some of the configurations proposed, we now assess the statistical significance of the improvement. For that analysis, we resort to the nonparametric Wilcoxon signed-rank test [41] with a significance value of $p < 0.05$.

Since the idea to compare whether the proposed architectures improve the base neural model, the analysis considers that each result obtained for each fold of the two corpora constitutes a sample of the distributions to be compared. It must be noted that, in this significance test, we focus on the symbol error rate as the metric to be analyzed.

Considering this assessment scheme, the results obtained are reported in Table 5.

Table 5. Statistical improvement over the base neural model of the different proposed schemes considering the Wilcoxon signed-rank test with a significance value of $p < 0.05$ for the symbol error rate metric. The symbols = and > represent that the error of the method in the row (the base neural architecture) is not significantly different or is significantly greater than that of the column, respectively.

	PreRNN		**InterRNN**		**PostRNN**	
	Raw	**Pretrained**	**Raw**	**Pretrained**	**Raw**	**Pretrained**
Baseline	=	=	>	>	>	>

A statistical assessment of the results obtained shows two clear conclusions, which were already intuitive from the previous analysis. On the one hand, the performance of the *PreRNN* architectures does not significantly differ from the *baseline* model. On the other hand, both of the *InterRNN* and *PostRNN* models do significantly improve the results of the *baseline* scheme since, as it can be seen, the error rate is consistently superior in the latter model than in the hybrid schemes.

As a last point to mention, the *InterRNN* and *PostRNN* models were also confronted with the Wilcoxon signed-rank test in the same conditions as in the previous analysis. Nevertheless, as expected from the average and deviation error rate figures aforementioned, there was no statistical difference between them.

6. Discussion

Current state-of-the-art OMR technologies, which are based on Convolutional Recurrent Neural Networks (CRNN), typically follow an end-to-end approach that operates at the staff level: they map the series of symbols that appear in an image of a single staff to a sequence of music symbol labels. Most commonly, these methods consider an agnostic music representation that defines every symbol as a 2-tuple element that encodes its two graphic components: shape and height. However, as mentioned in Section 1, holistic OMR systems [12,16,20] do not take advantage of this double dimension since each possible combination of shape and height is represented as a unique category. This leads to a recognition formulation completely equivalent to the one used in fields as text recognition [21]; hence, the particularities of this data and notation are neglected in the framework.

Nevertheless, we consider that exploitation of this trademark of music notation could increase the recognition rates, as recent work have provided insights about its benefits [22–25]. Considering the end-to-end neural-based framework by Calvo-Zaragoza et al. [20] as a starting point, we pose the following hypothesis: devoting independent CRNN schemes to separately exploit the two aforementioned graphical components of agnostic labeling and, in time, gathering them at the actual neural level might boost the overall performance of the model. In a practical sense, the considered CRNN architectures for each graphical component are equivalent to that of the starting point to perform a fair performance comparison between both the existing and proposed approaches.

In this work, we empirically evaluated three different integration policies differing in the point in which the specialized branches are merged: the *PreRNN*, which merges the information after the convolutional stage; the *InterRNN*, which joins the different branches after a first recurrent layer; and the *PostRNN*, which gathers the information after a second recurrent stage. These architectures are evaluated on two collections of monophonic early music manuscripts, namely the *Capitan* and the *Il Lauro Secco* corpora. The results presented in Section 5 show that the merging point affects the performance, as both the *InterRNN* and *PostRNN* models significantly improve the results of the baseline considered. Moreover, the *InterRNN* model yields the best overall performance, reducing the baseline error rate by 25.6% and 14.4% for the *Capitan* and *Il Lauro Secco* corpora, respectively.

In spite of this benefit, the proposed framework entails an increase in the network complexity with respect to the base model considered. Nevertheless, this increase does not imply more training data as each part of the network is meant to specialize in certain features of the staff image. Moreover, regarding processing time, the increase in the complexity of the network does not affect the inference phase, as the different branches work concurrently. The main drawback in this case though is the training lapse, which is certainly expanded. However, note that this drawback is assumable given the remarkable error rate decrease with respect to the figures obtained with the baseline considered.

Finally, let us note the generalization capability of the proposal, regarding the reach and performance of this neural architecture when tackling different corpora. As mentioned in Section 1, CRNN-based models are rather adaptable since the same architecture is capable of obtaining very competitive recognition rates on two very different corpora. Therefore, the inherent lack of generalization of heuristic systems mentioned in Section 1 is considered solved under the CRNN paradigm.

7. Conclusions

Holistic symbol recognition approaches have proven their usefulness in the context of Optical Music Recognition (OMR) since, given that no alignment is required between the input score and the sequence of output elements, corpora are relatively easy to create. In such a context, these sets are commonly labeled using either a *semantic* notation, which codifies the actual musical meaning of each element in the score, or an *agnostic* representation, which encodes the elements as a combination of a shape tag and the vertical position (height) in the score. While this latter representation lacks the underlying music sense of the former, it has the clear advantage of perfectly suiting an image-based symbol recognition task as the vocabulary is defined directly on visual information. However, in spite of having been considered for a number of work, it is still unclear how to properly take advantage of the fact that each symbol is actually a combination of two individual primitives representing the shape of the element and its height.

This work presents an end-to-end approach that exploits this two-dimensional nature of the agnostic music notation to solve the OMR task at a staff-line level. We considered two Convolutional Recurrent Neural Network (CRNN) schemes to concurrently exploit the shape and height information and to then merge them at the actual neural level. Three different integration policies were empirically studied: (i) the *PreRNN* one, which joins the shape and height features right before the recurrent block; (ii) the *InterRNN* one, which does the merging process after the first recurrent layer; and (iii) the *PostRNN* one, which collects both sources of information after the recurrent block. The results obtained confirm that the gathering point impacts the performance of the model, with the *InterRNN* and *PostRNN* models being the ones that significantly decrease the error rate with respect to the baseline considered. Quantitatively, in the best-case scenarios, this error reduction ranges between 14.4% and 25.6% referred to the base neural model.

In light of the obtained conclusions, this work opens new research points to address. In that sense, future work considers extending the approach from this monophonic context to a homophonic one [42]. Additionally, given the variability of music notation and the relative scarcity of existing labeled data, we aim to explore *transfer learning* and *domain adap-*

tation techniques to study different strategies to properly exploit the knowledge gathered from a given corpus on a different one. Moreover, we also consider that other network architectures may provide some additional insights to the ones obtained in this work as well as more competitive recognition rates. Finally, since the *greedy* decoding strategy considered does not take advantage of the actual Language Model inferred in the recurrent layer of the network, our premise is that more sophisticated decoding policies may report improvements in the overall performance of the proposal.

Author Contributions: Conceptualization, M.A.-C. and J.J.V.-M.; methodology, M.A.-C. and J.J.V.-M.; software, M.A.-C. and J.J.V.-M.; validation, M.A.-C. and J.J.V.-M.; formal analysis, M.A.-C. and J.J.V.-M.; writing, M.A.-C. and J.J.V.-M. All authors have read and agreed to the published version of the manuscript.

Funding: This research work was partially funded by the University of Alicante through project GRE19-04, by the "Programa I+D+i de la Generalitat Valenciana" through grant APOSTD/2020/256, and by the Spanish Ministerio de Universidades through grant FPU19/04957.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data considered for this work is already available from their respective cited sources. No new data was generated in this work.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BLSTM	Bidirectional Long Short-Term Memory
CRNN	Convolutional Recurrent Neural Network
CTC	Connectionist Temporal Classification
OMR	Optical Music Recognition
ReLU	Rectified Linear Unit
seq2seq	Sequence-to-Sequence
Seq-ER	Sequence Error Rate
Sym-ER	Symbol Error Rate

References

1. Schedl, M.; Gómez Gutiérrez, E.; Urbano, J. Music information retrieval: Recent developments and applications. *Found. Trends Inf. Retr.* **2014**, *8*, 127–261. [CrossRef]
2. Treitler, L. The early history of music writing in the west. *J. Am. Musicol. Soc.* **1982**, *35*, 237–279. [CrossRef]
3. Strayer, H.R. From Neumes to Notes: The Evolution of Music Notation. *Music. Offer.* **2013**, *4*, 1–14. [CrossRef]
4. Pugin, L. The challenge of data in digital musicology. *Front. Digit. Humanit.* **2015**, *2*, 4. [CrossRef]
5. Jones, G.; Ong, B.; Bruno, I.; Kia, N. Optical music imaging: Music document digitisation, recognition, evaluation, and restoration. In *Interactive Multimedia MUSIC technologies*; IGI Global: Hershey, PA, USA, 2008; pp. 50–79.
6. Baró, A.; Riba, P.; Calvo-Zaragoza, J.; Fornés, A. From optical music recognition to handwritten music recognition: A baseline. *Pattern Recognit. Lett.* **2019**, *123*, 1–8. [CrossRef]
7. Calvo-Zaragoza, J.; Hajič, J., Jr.; Pacha, A. Understanding optical music recognition. *ACM Comput. Surv. (CSUR)* **2020**, *53*, 1–35. [CrossRef]
8. Baró, A.; Badal, C.; Fornés, A. Handwritten Historical Music Recognition by Sequence-to-Sequence with Attention Mechanism. In Proceedings of the 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR), Dortmund, Germany, 8–10 September 2020; pp. 205–210.
9. Bainbridge, D.; Bell, T. The Challenge of Optical Music Recognition. *Comput. Humanit.* **2001**, *35*, 95–121. [CrossRef]
10. Rebelo, A.; Fujinaga, I.; Paszkiewicz, F.; Marçal, A.; Guedes, C.; Cardoso, J. Optical music recognition: State-of-the-art and open issues. *Int. J. Multimed. Inf. Retr.* **2012**, *1*. [CrossRef]
11. Rebelo, A.; Capela, G.; Cardoso, J.S. Optical recognition of music symbols. *Int. J. Doc. Anal. Recognit. (IJDAR)* **2010**, *13*, 19–31. [CrossRef]
12. Calvo-Zaragoza, J.; Rizo, D. End-to-end neural optical music recognition of monophonic scores. *Appl. Sci.* **2018**, *8*, 606. [CrossRef]
13. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]

14. O'Mahony, N.; Campbell, S.; Carvalho, A.; Harapanahalli, S.; Hernandez, G.V.; Krpalkova, L.; Riordan, D.; Walsh, J. Deep learning vs. traditional computer vision. In *Science and Information Conference*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 128–144.
15. Castro, F.M.; Marín-Jiménez, M.J.; Guil, N.; Schmid, C.; Alahari, K. End-to-end incremental learning. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 233–248.
16. Calvo-Zaragoza, J.; Valero-Mas, J.J.; Pertusa, A. End-To-End Optical Music Recognition using Neural Networks. In Proceedings of the ISMIR, Suzhou, China, 23–27 October 2017; pp. 472–477.
17. Pugin, L. Optical Music Recognitoin of Early Typographic Prints using Hidden Markov Models. In Proceedings of the ISMIR, Victoria, BC, Canada, 8–12 October 2006; pp. 53–56.
18. Calvo-Zaragoza, J.; Toselli, A.H.; Vidal, E. Early handwritten music recognition with hidden markov models. In Proceedings of the 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), Shenzhen, China, 23–26 October 2016; pp. 319–324.
19. Rizo, D.; Calvo-Zaragoza, J.; Iñesta, J.M.; Fujinaga, I. About agnostic representation of musical documents for Optical Music Recognition. In Proceedings of the Music Encoding Conference, Tours, France, 16–17 May 2017.
20. Calvo-Zaragoza, J.; Toselli, A.H.; Vidal, E. Handwritten music recognition for mensural notation with convolutional recurrent neural networks. *Pattern Recognit. Lett.* **2019**, *128*, 115–121. [CrossRef]
21. Castellanos, F.J.; Calvo-Zaragoza, J.; Inesta, J.M. A neural approach for full-page optical music recognition of mensural documents. In Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR, Montreal, QC, Canada, 12–16 October 2020; pp. 23–27.
22. van der Wel, E.; Ullrich, K. Optical Music Recognition with Convolutional Sequence-to-Sequence Models. In Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, 23–27 October 2017; pp. 731–737.
23. Nuñez-Alcover, A.; de León, P.J.P.; Calvo-Zaragoza, J. Glyph and position classification of music symbols in early music manuscripts. *Iberian Conference on Pattern Recognition and Image Analysis*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 159–168.
24. Villarreal, M.; Sánchez, J.A. Handwritten Music Recognition Improvement through Language Model Re-interpretation for Mensural Notation. In Proceedings of the 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR), Dortmund, Germany, 8–10 September 2020; pp. 199–204. [CrossRef]
25. Ríos-Vila, A.; Calvo-Zaragoza, J.; Inesta, J.M. Exploring the two-dimensional nature of music notation for score recognition with end-to-end approaches. In Proceedings of the 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR), Dortmund, Germany, 8–10 September 2020; pp. 193–198.
26. van der Wel, E.; Ullrich, K. Optical Music Recognition with Convolutional Sequence-to-Sequence Models. *arXiv* **2017**, arXiv:1707.04877.
27. Shatri, E.; Fazekas, G. Optical Music Recognition: State of the Art and Major Challenges. *arXiv* **2020**, arXiv:2006.07885.
28. Shi, B.; Bai, X.; Yao, C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2298–2304. [CrossRef]
29. Liu, A.; Zhang, L.; Mei, Y.; Lian, S.; Han, M.; Cheng, W.; Liu, Y.; Cai, Z.; Zhu, Z.; Han, B.; et al. Residual Recurrent CRNN for End-to-End Optical Music Recognition on Monophonic Scores. *arXiv* **2020**, arXiv:2010.13418.
30. Wick, C.; Puppe, F. Experiments and detailed error-analysis of automatic square notation transcription of medieval music manuscripts using CNN/LSTM-networks and a neume dictionary. *J. New Music. Res.* **2021**, *50*, 18–36. [CrossRef]
31. Schneider, D.; Korfhage, N.; Mühling, M.; Lüttig, P.; Freisleben, B. Automatic Transcription of Organ Tablature Music Notation with Deep Neural Networks. *Trans. Int. Soc. Music. Inf. Retr.* **2021**, *4*, 14–28. [CrossRef]
32. Rizo, D.; Calvo-Zaragoza, J.; Iñesta, J.M. Muret: A music recognition, encoding, and transcription tool. In Proceedings of the 5th International Conference on Digital Libraries for Musicology, Paris, France, 23–27 September 2018; pp. 52–56.
33. Calvo-Zaragoza, J.; Vigliensoni, G.; Fujinaga, I. One-Step Detection of Background, Staff Lines, and Symbols in Medieval Music Manuscripts with Convolutional Neural Networks. In Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, 23–27 October 2017; Cunningham, S.J., Duan, Z., Hu, X., Turnbull, D., Eds.; pp. 724–730.
34. Graves, A. Supervised sequence labelling. In *Supervised Sequence Labelling with Recurrent Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 5–13.
35. Graves, A.; Fernández, S.; Gomez, F.; Schmidhuber, J. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In Proceedings of the 23rd International Conference on Machine Learning, ICML'06, Corvallis, OR, USA, 25–29 June 2006; ACM: New York, NY, USA, 2006; pp. 369–376.
36. Zenkel, T.; Sanabria, R.; Metze, F.; Niehues, J.; Sperber, M.; Stüker, S.; Waibel, A. Comparison of Decoding Strategies for CTC Acoustic Models. *Proc. Interspeech 2017* **2017**, 513–517. [CrossRef]
37. Calvo-Zaragoza, J.; Toselli, A.H.; Vidal, E. Handwritten music recognition for mensural notation: Formulation, data and baseline results. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; Volume 1, pp. 1081–1086.

38. Parada-Cabaleiro, E.; Batliner, A.; Schuller, B.W. A Diplomatic Edition of Il Lauro Secco: Ground Truth for OMR of White Mensural Notation. In Proceedings of the ISMIR, Delft, The Netherlands, 4–8 November 2019; pp. 557–564.
39. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
40. Levenshtein, V.I. Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl.* **1966**, *10*, 707–710.
41. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
42. Alfaro-Contreras, M.; Calvo-Zaragoza, J.; Iñesta, J.M. Approaching end-to-end optical music recognition for homophonic scores. In Proceedings of the Iberian Conference on Pattern Recognition and Image Analysis, Madrid, Spain, 1–4 July 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 147–158.