Expert Systems With Applications 209 (2022) 118211

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Region-based layout analysis of music score images

Francisco J. Castellanos [*], Carlos Garrido-Munoz, Antonio Ríos-Vila, Jorge Calvo-Zaragoza

Department of Software and Computing Systems, University of Alicante, Carretera San Vicente del Raspeig s/n, 03690, Alicante, Spain

ARTICLE INFO

Keywords:
Optical Music Recognition
Layout Analysis
Image augmentation
Object detection

ABSTRACT

The Layout Analysis (LA) stage is of vital importance to the correct performance of an Optical Music Recognition (OMR) system. It identifies the regions of interest, such as staves or lyrics, which must then be processed in order to transcribe their content. Despite the existence of modern approaches based on deep learning, an exhaustive study of LA in OMR has not yet been carried out with regard to the performance of different models, their generalization to different domains or, more importantly, their impact on subsequent stages of the pipeline. This work focuses on filling this gap in the literature by means of an experimental study of different neural architectures, music document types, and evaluation scenarios. The need for training data has also led to a proposal for a new semi-synthetic data-generation technique that enables the efficient applicability of LA approaches in real scenarios. Our results show that: (i) the choice of the model and its performance are crucial for the entire transcription process; (ii) the metrics commonly used to evaluate the LA stage do not always correlate with the final performance of the OMR system, and (iii) the proposed data-generation technique enables state-of-the-art results to be achieved with a limited set of labeled data.

1. Introduction

The digitization of music manuscripts helps preserve and disseminate this valuable heritage. However, simply obtaining a digital image from the original source is not sufficient to enable the computational use of this material, and it is, therefore, necessary to transcribe the content into a digital format.

The manual transcription of music sources is a time-consuming task. Given the countless number of music manuscripts scattered around the world, there are ongoing efforts to automate this process by means of artificial intelligence. The research field that studies how to automatically transcribe music notation from scanned documents into a structured digital format is denominated as Optical Music Recognition (OMR) (Calvo-Zaragoza et al., 2020). The structural complexity of music notation, along with the great variability as regards writing styles, engraving mechanisms or types of notation – such as neumatic, mensural or modern Western notation – makes the OMR challenge far from straightforward.

Music manuscripts may contain not only music but also text, lyrics or document metadata, and it is for this reason that several tasks in the traditional OMR workflow focus on the document itself with respect to the distribution of its different regions (Rebelo et al., 2012). This is usually referred to as LA, which is also common in other document contexts such as text recognition (Binmakhashen & Mahmoud, 2019). The main purpose of LA is to identify the relevant information from the whole image, thus facilitating the eventual objective of these systems: the transcription of their content. This can be done using two broad possible strategies LA: *pixel-wise*, in which each pixel is explicitly classified according to a set of different categories or information layers contained in images, such as staff lines, text, music notation, etc.; or *region-based*, in which a list of polygons (typically bounding boxes) defines the regions in which the elements are located. This work is focused on the latter approach, i.e., the region-based approach.

Common formulations consider region-based LA as an object detection task, in which each meaningful element or region of interest is located and classified into a set of predefined categories (Liu et al., 2020). Despite its importance in the OMR context, and although the literature contains a large number of general-purpose object-detection approaches, no comprehensive study has been performed in order to assess the behavior of different models for LA when applied to music scores, unlike that which has occurred in other research fields, such as Handwritten Text Recognition (Studer et al., 2019; Zhong et al., 2019). In this work, we aim to fill this gap in the literature by performing thorough experiments with several object-detection approaches that are evaluated in different scenarios.

Furthermore, the state of the art as regards LA involves the use of machine learning, and particularly deep learning techniques (LeCun et al., 2015). The excellent performance demonstrated in several

[*] Corresponding author.
E-mail addresses: fcastellanos@dlsi.ua.es (F.J. Castellanos), carlos.garrido@ua.es (C. Garrido-Munoz), arios@dlsi.ua.es (A. Ríos-Vila), jcalvo@dlsi.ua.es (J. Calvo-Zaragoza).

https://doi.org/10.1016/j.eswa.2022.118211
Received 25 January 2022; Received in revised form 28 June 2022; Accepted 16 July 2022
Available online 21 July 2022

computer vision contexts makes this type of techniques appropriate for the task discussed herein. However, the application of these techniques to LA is not yet straightforward in real-world scenarios, since one of the major challenges is the need for sufficiently representative ground-truth data. This issue is particularly relevant in the context of music documents, given that manuscripts are highly heterogeneous and that very few manuscripts have been properly annotated. Reusing data from different collections to process another is, therefore, reasonably assumed to be ineffective. We propose to address this common issue by using an algorithm with which to generate semi-synthetic images in order to increment the set of available annotated data to be used as a reference for the deep learning approach.

Moreover, it should be noted that no specific metric has been designed for OMR in order to evaluate the performance of LA. One of the most common metrics used for assessment in the object-detection field is the mean Average Precision (mAP) metric, which has also been employed in the context of music (Jia et al., 2021; Pacha et al., 2018). However, no analysis has yet been carried out regarding whether it is indeed an appropriate metric that correlates with the quality of the bounding boxes extracted in order to eventually transcribe music. We, therefore, discuss the results obtained with different metrics in order to be able to state which is the most suitable for LA when applied in OMR.

Finally, since LA is one of the earliest steps in the OMR workflow, any inaccuracies might be critical for the eventual transcription, signifying that it is crucial to study its influence on the whole process. It is essential to analyze the interaction between LA and the eventual transcription in order to discover the most appropriate way in which to address the OMR process. However, LA has usually been evaluated as an individual task without properly analyzing this question, and no existing study covers it.

To summarize, the contribution of this paper can be divided into the following points:

- Carrying out the first comprehensive study of object-detection models for LA in music score images.
- Analyzing and discussing the correlation between the common metrics used in object detection and the quality of the bounding boxes retrieved in the LA process for OMR.
- Proposing a new semi-synthetic data-generation method for LA, in addition to carrying out a thorough study of its usefulness.
- A goal-directed analysis of the influence of LA on the eventual transcription.

The remainder of this paper is organized as follows: the state of the art of LA is detailed in Section 2, while the different architectures considered and our data-augmentation mechanism for LA are described in Section 3. The experimental setup, along with the corpora and metrics considered, are explained in Section 4, and Section 5 shows the results obtained after carrying out staff-retrieval and end-to-end recognition experiments, in addition to the corresponding analysis and discussion of them. Finally, the main conclusions of the work are summarized in Section 6.

2. Background

Two main perspectives of LA can, in broad terms, be considered in the context of OMR: processing the document at pixel level or at region level.

The former perspective was traditionally addressed through the use of different strategies that can be found in literature. Before deep learning techniques were applied, other conventional systems were employed by means of heuristic techniques. For example, in order to separately extract the staff lines and lyrics from sheet music, Burgoyne et al. (2009) proposed a heuristic method based on the Hough transform that could be used to detect the waved text and staff lines. Although the staff lines are highly necessary as regards recognizing the pitch of the symbols, many OMR workflows are based on this process, which is employed to perform a connected component analysis of the remaining music notation. There is a review that shows the earlier methods used for staff removal (Dalitz et al., 2008), but new techniques have also been developed in order to address this question through the use of heuristic methods (Géraud, 2014; dos Santos Cardoso et al., 2009). Of the topics related to OMR, there is, among others, a review (Rebelo et al., 2012) that gathers this type of solutions together in order to perform LA.

Despite the fact that these heuristic strategies may obtain good results in controlled scenarios, they are poorly generalizable, signifying that they are not, in practice, suitable for the processing of scanned documents. The major challenge in this respect is the great variability of this type of images owing to multiple factors, e.g. the degree of degradation, contrast, the color of the ink employed or skew variations, thus making it a difficult task to perform. The main focus as regards obtaining more generalizable models has been machine learning techniques, and particularly deep learning techniques. For example, Calvo-Zaragoza et al. (2018) presented a Convolutional Neural Network (CNN)-based architecture with which to perform LA by classifying each pixel of the image according to a set of categories. However, the method takes a long time because it processes each pixel in the image. An attempt has been made to address this time-consuming issue by employing an image-to-image strategy (Castellanos et al., 2018), which is based on a series of encoder–decoder architectures (the so-called SAE) and trains each one in order to extract a particular information layer.

The region-wise perspective can, meanwhile, be considered as an object-detection process in which the objects are different parts of the document, such as staff regions or lyrics. Several previous works have followed this approach. One of the first was that of Bosch et al. (2016), which used Hidden Markov Models to carry out LA in order to extract text and staff regions from music score images, while Quirós et al. (2019) proposed the use of an Artificial Neural Network architecture to extract the different regions of interest from a music document. In their work, Pacha (2019) developed an incremental method for the training of supervised models with a combination of annotated and predicted images so as to extract the bounding boxes of staves. Moreover, Waloschek et al. (2019) proposed a neural network approach that could be used to extract the bounding boxes of the system measures from a music score image, focusing on the alignment between them. A full-page framework based on two steps – staff recognition and end-to-end transcription – was recently proposed (Castellanos et al., 2020). With respect to the first step, it performed LA in order to extract the staff regions by means of an SAE and connected-component analysis. In their work, meanwhile, Kletz and Pacha (2021) proposed the use of Faster Region-based Convolutional Neural Network (Faster R-CNN) (Ren et al., 2015) to detect the bounding boxes of staves and measures.

In addition, note that a large number of region-based LA methods use mAP as a metric to evaluate the quality of the bounding boxes (Huang et al., 2019; Kletz & Pacha, 2021; Pacha et al., 2018; Waloschek et al., 2019), but no study of the suitability of this metric for OMR can be found in the literature.

3. Methodology

This section provides a description of the methodology considered for LA. It is divided into two parts: the description of the different object-detection models deemed appropriate for this task, and the definition of our data augmentation proposal with which to generate new semi-synthetic images, which is particularly useful when there is insufficient annotated data.

3.1. Object-detection architectures for Layout Analysis

We considered several well-known general-purpose models for the task of applying object detection in LA. These were selected owing to their popularity and considerable capabilities in multiple areas, in addition to the fact that they cover various neural strategies, such as one-stage or two-stage models or even pixel-wise segmentation. We specifically used those shown below:

- Faster R-CNN (Ren et al., 2015) is a two-stage detection model that includes a Region Proposal Network (RPN) to Fast R-CNN (Girshick, 2015). This model uses the last convolutional layer of the backbone as a feature map and attempts to extract proposals for classification and localization directly through the RPN. These proposals are also used to train the classifier, enabling it to create a unified network. Since convolutional features are shared, the efficiency of training increases when compared to other previous architectures such as Fast R-CNN and R-CNN. This model has proven to be highly efficient and to perform well in several scenarios, thus making it an ideal candidate for carrying out LA. We consider this detector alongside the ResNet50 backbone (He et al., 2016).
- RetinaNet (Lin et al., 2020) is a one-stage object-detection model composed of a backbone and two sub-networks. The backbone part computes the convolutional feature map over the input image, typically relying on ResNet and adopting the Feature Pyramid Network (FPN) in order to extract proposals. The sub-network part is composed of classification and box regression networks. RetinaNet attempts to solve the common problem of imbalanced data, in which there are different numbers of samples for each class. This issue usually causes a bias in the training by tending to predict the majority class (usually, “background”). RetinaNet addresses this by means of a focal loss function, which dynamically shifts weights in order to decrease the contribution of well-classified samples and focuses on misclassifications by means of the modulating factor of the focal loss. This model is especially interesting for LA, since the content of a music document is very varied and may contain disparate elements, such as a different number of staff and text regions. We use this detector in combination with ResNet50 and FPN.
- SSD (Liu et al., 2016) is a one-stage model that takes feature maps in order to generate multi-scale proposal predictions. It retrieves objects in one step and explicitly divides the predictions by employing an aspect ratio. We use this detector with VGG16 (Simonyan & Zisserman, 2015) as a backbone, since it is the basis of the original work.
- SAE is a Fully-Convolutional Network (FCN), and specifically a U-net architecture (Ronneberger et al., 2015), which is able to classify each pixel in an image according to a set of categories. This type of architecture is composed of two parts: an encoder that extracts the relevant features with combinations of convolutional and pooling layers, and a decoder that inverts the encoder operation with convolutional and up-sampling layers until the size of the input image is retrieved. The SAE model provides a probabilistic map whose elements contain the probability of each pixel belonging to a specific class. This model has been successfully used for the staff-retrieval task with music score images (Castellanos et al., 2020). It is important to highlight that to be able to apply this method in LA, a post-process is required in order to convert the probabilistic map obtained by the neural network into a set of bounding boxes by performing a connected-component analysis.

It should be noted that the three first architectures presented–Faster R-CNN, RetinaNet and SSD–rely on the use of a handcrafted technique called Non-Maximum-Suppression (Neubeck & Van Gool, 2006) to deal with multiple detections of the same object.

Moreover, the SAE method has the restriction of being unable to detect overlapped bounding boxes. We have, therefore, followed the strategy of Castellanos et al. (2020) to vertically reduce the bounding boxes by 20% of the ground truth in order to mitigate this restriction. With regard to the predictions, after extracting the coordinates of the bounding boxes, the method vertically increases the predicted bounding boxes by the same ratio. Note that this alteration is not necessary for the other methods. In addition, SAE does not provide a confidence value for each precision, which is the degree of certainty that the model has for its estimation, although the other models do provide it.

3.2. Semi-synthetic data-generation for Layout Analysis

Let us consider $\mathcal{D}$, a collection of labeled images consisting of pairs $(\mathcal{I}, \mathcal{R})$, in which $\mathcal{I}$ is an image and $\mathcal{R}$ is its respective ground-truth bounding boxes or regions. In this work, global coordinates are used to define each of these regions with its location within $\mathcal{I}$ and the class to which it belongs—`staff` or `text`, but more classes could be applicable. The idea behind our data-augmentation algorithm is to take advantage of the often scarce ground-truth data available in the OMR context in order to build new semi-synthetic images composed of a combination of individual elements extracted from the original images.

Algorithm 1 Image-generation algorithm proposed.

```
1:  function Image-Generation($\mathcal{D}$, $n$, $\Phi$)
2:      $\mathcal{S} \leftarrow \emptyset$
3:      for $i \leftarrow 1$ to $n$ do
4:          $\mathcal{I}, \mathcal{R} \leftarrow$ image-selection-policy($\mathcal{D}$)
5:          $\mathcal{I}_s \leftarrow$ background-estimation($\mathcal{I}$)
6:          $\mathcal{R}_s \leftarrow \emptyset$
7:          for each $r \in \mathcal{R}$ do
8:              $r_s, i_s \leftarrow$ region-selection-policy($r$, $\mathcal{D}$)
9:              $x_r, y_r \leftarrow$ reference-global-coordinates($r$)
10:             $r_s \leftarrow$ update-global-coordinates($r_s$, $x_r$, $y_r$)
11:             $r_s, i_s \leftarrow$ distortion-policy($r_s$, $i_s$, $\Phi$)
12:             $\mathcal{R}_s \leftarrow \mathcal{R}_s \cup r_s$
13:             $\mathcal{C}_s \leftarrow$ ink-detection($i_s$)
14:             for each $(x_s, y_s) \in \mathcal{C}_s$ do
15:                 $\mathcal{I}_s[x_r + x_s][y_r + y_s] \leftarrow i_s[x_s][y_s]$
16:             end for
17:         end for
18:         $\mathcal{S} \leftarrow \mathcal{S} \cup (\mathcal{I}_s, \mathcal{R}_s)$
19:     end for
20:     return $\mathcal{S}$
21: end function
```

The proposed data-augmentation mechanism is described in Algorithm 1. The parameters received by this algorithm comprise the collection of images $\mathcal{D}$, the number of new images n that have to be generated, and the distortion policy Φ that configures the type of distortion to be applied. The principal idea is to build a new dataset $\mathcal{S}$ from $\mathcal{D}$ with n generated images. $\mathcal{S}$ consists of a set of pairs in the form of $(\mathcal{I}_s, \mathcal{R}_s)$, in which $\mathcal{I}_s$ is a semi-synthetic image and $\mathcal{R}_s$ is the respective ground-truth data for the bounding boxes of $\mathcal{I}_s$.

In order to obtain a realistic image $\mathcal{I}_s$, the algorithm first selects an existing image $\mathcal{I}$ and its respective ground-truth data $\mathcal{R}$ from $\mathcal{D}$ by means of the function *image-selection-policy*$(\cdot, \cdot)$, as shown in line 4. In this work, this selection was carried out randomly, but other policies could also be applied.

The next step consists of generating a background for the new image. This is performed in line 5 by the function *background-estimation*$(\cdot)$, which applies a process to the selected image $\mathcal{I}$ in order to build the basis of the new image. We propose using a blurring operation, with the objective of fading the content of the image and obtaining an empty image with a similar background to that of the original one. We applied a Gaussian blur operation with a high kernel value (half of the

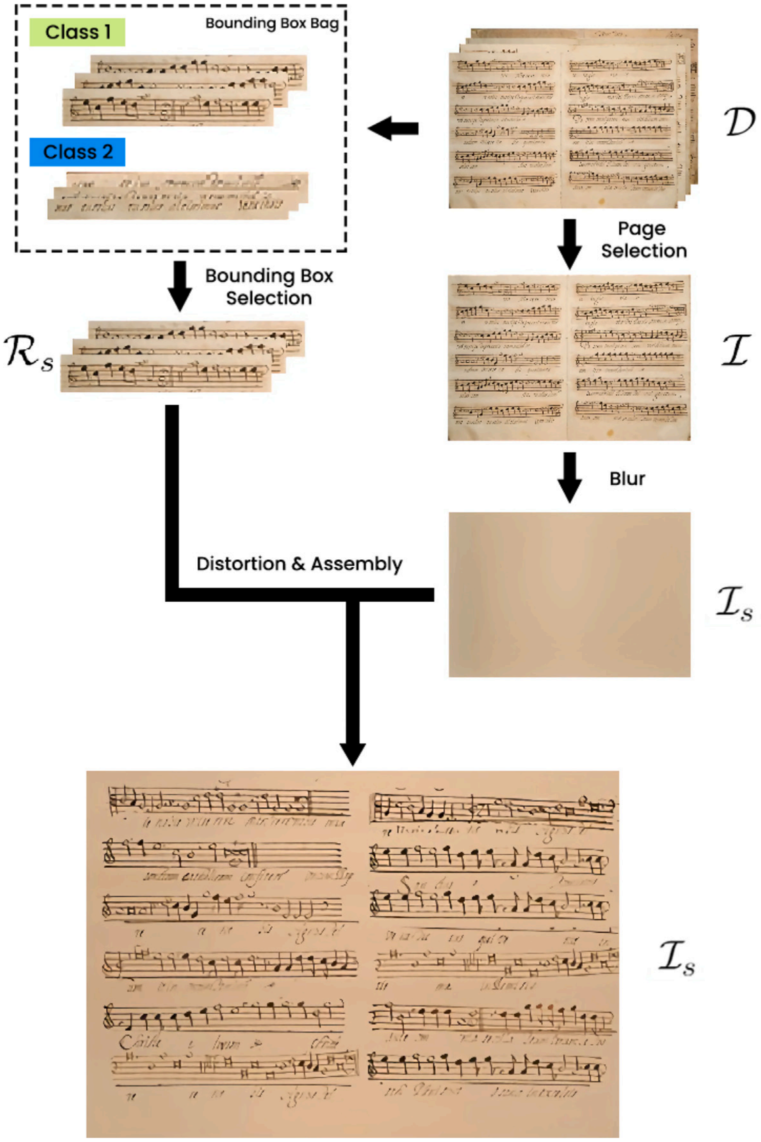


Fig. 1. Overview of the data-generation algorithm proposed.

image size in our case). This usually erases all possible traces from ink information shown in these documents, as shown in Fig. 1).

Once the background has been built, and in order to keep the structure of the original image $\mathcal{I}$, we propose replacing each region $r \in \mathcal{R}$ with another same-class region r_s that is available in $\mathcal{D}$. For example, if r represents a bounding box of a music staff, the new region r_s will also be a staff region selected from all those available in $\mathcal{D}$ in order to then locate them in the same position as r. This selection is carried out in line 8 by the function *region-selection-policy*$(\cdot,\cdot)$, which applies a selection policy in order to search for a replacement for each $r \in \mathcal{R}$. In our case, we consider a random selection of same-class regions. Note that, in addition to extracting the new region r_s, the portion of image i_s represented by the bounding box is also extracted.

Since r_s contains coordinates that are relative to the original image from which r_s was extracted, it is necessary to adjust the coordinates to those relative to the new image. The method, therefore, extracts the reference coordinates x_r, y_r from r by means of the function *reference-global-coordinates*$(\cdot)$ shown in line 9. In our case, we use the upper-left corner of the bounding boxes as reference coordinates. These coordinates are then used to update the coordinates of the new bounding box r_s by means of the function *update-global-coordinates*$(\cdot,\cdot,\cdot)$, which is in line 10.

For the sake of more variability, in line 11, the function *distortion-policy*$(\cdot,\cdot,\cdot)$ applies a distortion policy Φ to i_s as an image-augmentation process. We apply a slight random rotation with respect to the center of each region. This rotation is of between −3° and 3° – a range used in previous work (López-Gutiérrez et al., 2021) for data augmentation in OMR – with respect to the original skew, and this is the same value for all the regions on a page, but is different for other pages. It should be noted that excessive rotation could lead to the overlapping of multiple bounding boxes, which would lead to the attainment of unrealistic images. This function also updates the coordinates of r_s according to the distortion applied. r_s must subsequently be included in $\mathcal{R}_s$, as stated in line 12, and, therefore, $r_s \in \mathcal{R}_s$.

At this point, the algorithm must dump the content of the new bounding box i_s onto the image generated $\mathcal{I}_s$. However, it is important to emphasize that although the background is obtained by processing a real image, it is not exactly the same as that in the original images. We, therefore, propose the use of only the relevant information from i_s – the pixels with ink – and avoiding the background pixels. In order to perform this, in line 13, the function represented as *ink-detection*$(\cdot)$ returns the set of coordinates C_s in the form (x_s, y_s), which represent the relative positions of the ink within i_s. Literature contains countless binarization approaches that can be used for this purpose (He

& Schomaker, 2019; Pastor-Pellicer et al., 2015). In this work, we have applied the well-known local-thresholding algorithm for binarization developed by Sauvola and Pietikäinen (2000), but any other could be used. The ink pixels in the i_s, which are indicated by the relative coordinates C_s, are then dumped onto $\mathcal{I}_s$ by using x_r, y_r to properly locate the ink information. This is performed in lines 14–16 of the algorithm.

Finally, once the above process has been completed for all the regions in $\mathcal{R}$, the new semi-synthetic image $\mathcal{I}_s$ and its respective ground-truth data r_s are included in $\mathcal{S}$, as stated in line 18, which contains the augmented dataset that will be returned at the end of the algorithm. The entire process is then repeated until n new images have been generated. The algorithm described is also shown schematically in Fig. 1.

It should also be noted that, although the documents belong to the same manuscript, the same-class regions of crossing pages may be of different sizes. We, therefore, considered skipping those replacements in which the inclusion of r_s within $\mathcal{R}_s$ causes overlapping between different bounding boxes or a part of the new region falls outside $\mathcal{I}_s$.

4. Experimental setup

This section covers the description of the corpora and metrics considered, along with the configuration of each model employed for this study. For the experimentation, each corpus was divided into three partitions: a training set, a validation set, and a test set. The training set is always fixed to 64 pages for all corpora (to ensure similar conditions in all of them), while the validation and test sets are equally divided with the remaining pages—which vary depending on the corpora (see Table 1 below). Note that the images from the testing set were not employed in the training process. This configuration was maintained throughout the experimentation, including the three evaluation scenarios studied in this work. Moreover, all the models considered in this experimentation were trained through 300 epochs with an early stopping of 30 epochs without improvement to the validation. We used a stochastic gradient descent (Bottou, 2010) optimizer with a learning rate of 0.001 and a momentum of 0.9. More information on the configuration of each model can be found in Appendix A.[1] Note also that we considered the generation of 100 semi-synthetic images with our data-generation method (see Algorithm 1) in our experimentation, but there is no restriction on the quantity of images that can be generated in practice.

4.1. Corpora

Several music corpora were considered for the experimentation. These were selected because of their dissimilar nature, as depicted in Fig. 2, in order to attain a better understanding of the generalization of the proposed methodology. We additionally selected collections containing the proper annotations for not only the layout analysis stage but also an eventual music transcription with an end-to-end approach. This was necessary in order to assess the impact of the layout analysis process in subsequent stages. We specifically considered the following corpora, whose details are shown in Table 1:

- Seils: This dataset contains 150 typeset pages of the Il Lauro Secco manuscript (Parada-Cabaleiro et al., 2019) corresponding to an anthology of 16th-century Italian madrigals in mensural notation.
- Capitan: This corpus is a compilation of 17th and 18th century manuscripts from the 'Cathedral of Our Lady of the Pillar' in Zaragoza (Spain).[2] This dataset is an evolution of the 'Zaragoza' corpus, which was created manually and introduced by Calvo-Zaragoza et al. (2016).
- FMT: The 'Fondo de Música Tradicional IMF-CSIC' corpus (Ros-Fábregas, 2021) consists of a collection of four groups of handwritten score sheets for popular Spanish songs transcribed by musicologists between 1944 and 1960. As it contains various manuscripts with dissimilar features, such as page color, image resolution or staff-region size, among others, these manuscripts have been clustered by similarity into two datasets: FMT-M and FMT-C, whose graphic differences are depicted in Figs. 2(c) and 2(d), respectively.

Table 1
Description of the corpora. The "Descr." column represents the description. Moreover, in the "Engraving" row, "Hw." signifies handwritten pages, whereas "Pr." represents printed pages.

Descr.	Seils	Capitan	FMT-M	FMT-C
Engraving	Pr.	Hw.	Hw.	Hw.
Pages	150	96	703	140
Lyrics	2 237	695	1 241	452
Staves	1 430	775	1 508	1 435
Symbols	31 589	17 115	11 327	5 766

Please note that the regions in all the corpora considered in this work are represented as rectangular bounding boxes, regardless of the underlying skew of the pages.

Furthermore, when evaluating LA, we study the behavior of the object-detection models in situations in which a different number of annotated images is used for the training process. In order to perform the same experiments for all the corpora, we, therefore, fixed a maximum of 64 pages to train the models, increasing from 1 to 64 in powers of two. Note that the number indicates the real pages. The experiments with synthetic data generation always make use of 100 images generated through the use of Algorithm 1, which takes only the real training pages indicated. Although we have considered 100 generated images, there is no limitation to the number of synthetic pages, since all regions are randomly selected and rotated in order to increase the variability of data. The other real pages were divided equally into validation and testing partitions.

4.2. Metrics

The proposed methodology was evaluated by considering different metrics, according to the specific experiment being carried out.

With regard to the LA experiments, we considered the COCO mAP metric (Lin et al., 2014), which is widely used to evaluate object-detection models. This metric computes the area under the precision–recall curve, considering a range of values of Intersection over Union (IoU), ranging from 0.5 to 0.95 in intervals of 0.05.

However, the importance of LA in OMR lies mainly in the region retrieval and not so much in how well the predictions fit the ground-truth bounding boxes. This signifies that mAP is not an appropriate metric for LA, as will be discussed at greater length in Section 5.2. We shall, therefore, also evaluate LA in terms of precision P, recall R and the harmonic mean F-score (F_1), which are computed as follows:

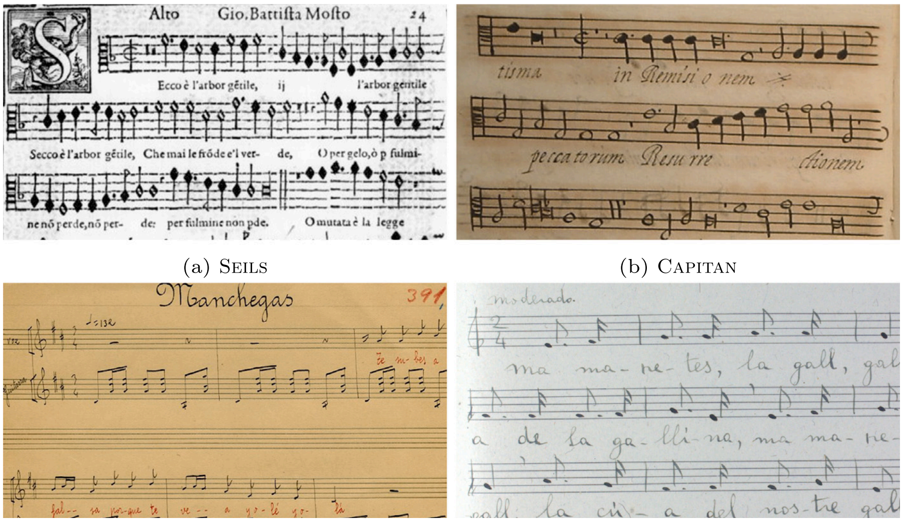
$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \quad \mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{1}$$

$$\mathrm{F}_1 = 2 \cdot \frac{\mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}}, \tag{2}$$

where TP, FP, and FN (in our context) represent *True Positives* or correctly classified regions, *False Positives* or type I errors refer to those predictions that do not match a real bounding box, and *False Negatives* or type II errors refer to the real regions that have not been detected, respectively. This matching between the predicted and the real bounding boxes is carried out in two steps. First, predictions are filtered by confidence value, except for the SAE (which does not provide such value). Then, these bounding boxes are assigned to a

[1] The code involved in the experimentation shown in this paper can be found in https://github.com/fjcastellanos/music_region_layout_analysis.

[2] RISM Code 'E-Zac' at https://rism.info/.



(a) Seils

(b) Capitan

(c) FMT-M

(d) FMT-C

Fig. 2. Samples of the corpora considered for the experimentation.

ground-truth box according to the IoU value and a threshold that is empirically studied. This is described in more detail in Section 5.2. Note that these metrics are evaluated and computed with respect to one class. For the evaluation of multiple classes using a single value, these metrics can instead be reformulated as the macro average, in which macro-precision, macro-recall and macro-F_1 – henceforth mP, mR and mF_1 are, respectively – the average of P, R and F_1 for all the classes involved.

We shall also evaluate the quality of the regions detected in terms of the ability of a state-of-the-art OMR model to retrieve the musical symbols from them. In this case, the effectiveness of the transcription system is typically measured using the Symbol Error Rate (SER) metric (Calvo-Zaragoza et al., 2019). Let H be the hypothesized sequence of music symbols and R be the ground-truth sequence, and let SER be computed by dividing the Levenshtein distance between H and R by the length of R. Note that, if a staff region is recognized as two or more regions, only the region with the greatest IoU is selected to be matched with the ground-truth staff. In this case, the other regions are considered to be FP.

5. Results

In this section, we analyze the results obtained after carrying out three evaluation scenarios. Various means were employed to analyze the performance or the goal of the experiment: (i) a standard evaluation, in which the typical metric in object detection (mAP) was used to assess the predicted bounding boxes; (ii) an evaluation in terms of retrieved regions, in which the estimations were measured by employing P, R and F_1, thus emphasizing the retrieval of bounding boxes rather than IoU; and finally, (iii) a goal-directed evaluation in which a study of the influence of the IoU and the confidence provided by the LA model in the final transcription – scored by means of SER – is discussed. Note that these metrics cannot, in practice, be used to evaluate the final results since annotations of the bounding boxes of the images are required in order to calculate the IoU of the predicted regions when compared to the real ones. They can, however, be used to validate the training of the models with a reference set (validation set) in order to determine whether or not the quality of the regions retrieved is better. They can also be calculated in order to optimize the results concerning the validation set in the training process.

5.1. Evaluation scenario I: Standard evaluation

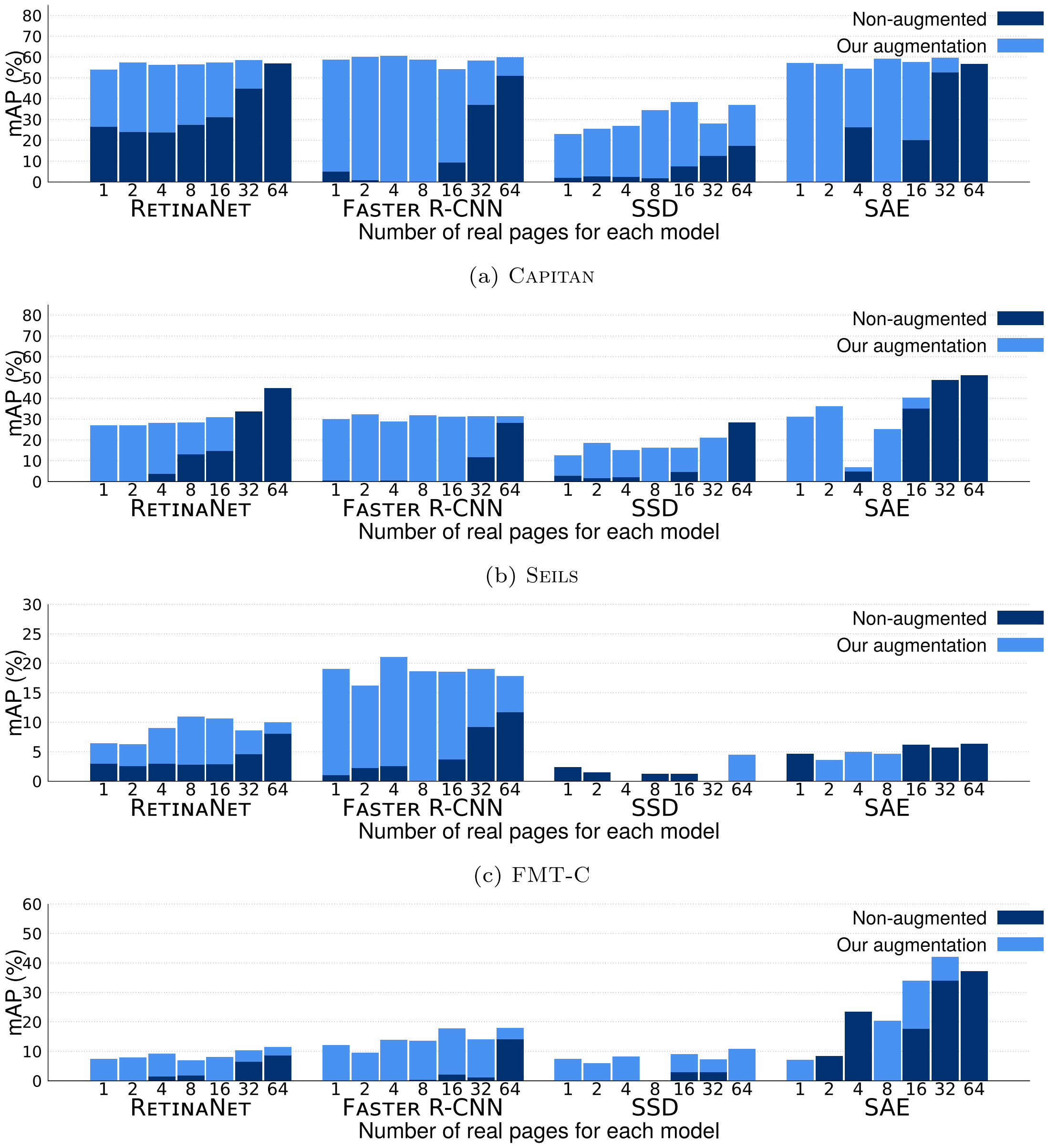
In this section, we present the results obtained after experimenting with a series of object-detection models whose purpose is to perform the LA process on music score images at the region level, as described in Section 3.1.

The models are evaluated in terms of COCO mAP in different situations of data availability with the aim of studying their behavior according to the number of annotated pages used to train them. Because of the cost of manually annotating music manuscripts, it is particularly relevant to analyze their behavior when limited annotated data is provided. We, therefore, also study the benefits of the algorithm proposed in Section 3.2 as regards building 100 semi-synthetic images and increasing the number of pages and variability of data. Fig. 3 provides a graphic representation of this metric in order to study the effectiveness of each model according to the number of real pages used to train them. It also shows the results obtained after the application of our data augmentation proposal when compared with the use of only the original images.

First note that the results are, in general, quite modest, mainly because of the rigorousness of the metric used. Moreover, although the amount of pages is crucial as regards optimizing mAP, a higher number of pages does not guarantee good results, depending on the difficulty of the corpus. Indeed, FMT attains more overlapping and a greater density of bounding boxes, especially in the case of FMT-C, which considerably increases the difficulty of the predictions and which translates into worse detection quality.

Despite this, it will be observed that, as expected, there is a similar trend for almost all the models, in which the fewer the number of actual pages, the worse the detection gets, since the models do not have sufficient reference data with which to learn patterns in order to generalize the detection. Nevertheless, when our data-augmentation algorithm is applied, these results are drastically improved, obtaining models that are more robust to the lack of ground-truth data.

RetinaNet and Faster R-CNN benefit most from data-augmentation, with improvements to all the datasets and nearly all the training data sizes, especially when a few real pages are available for training the models. The results show an increase in stability for these models, since they achieve more robustness, especially in those cases in which the data are limited. The results for SAE and SSD are also improved with our data-augmentation algorithm, but there are fewer cases in which the augmented data are better than the original ones, depending mainly on the corpus considered. These last models would appear to be more sensitive to the overlapping of the regions, since they are the models that attain the worst results as regards FMT-C. In this respect, Faster R-CNN with data augmentation obtains the best figures in this challenge corpus, although the results do not reach 25% of mAP in either case.



(a) Capitan

(b) Seils

(c) FMT-C

(d) FMT-M

Fig. 3. Results, in terms of COCO mAP(%), obtained for different object-detection models in different scenarios, in which the number of original documents available is scarce. The "Non-augmented" bars indicate the results obtained with only original images, whereas the "Our augmentation" bars represent the cases in which our data-generation algorithm is used to build 100 synthetic images.

It should be noted that deep neural networks generally require great amounts of data in order to train the models (Goodfellow et al., 2016). If there is insufficient data for training, the models may be unstable, leading to over-fitting and the attainment of unexpected results. For example, in the case of Capitan, note that SAE obtains about 25% of mAP with 4 real pages and about 20% when 16 real pages are used. There is no doubt that these metric values of less than 25% support the idea that there is a need for larger training sets with which to build usable models for this field. Almost all the experiments show this phenomenon when a few real pages are available, and when our augmentation algorithm is applied to generate 100 augmented images, the model undergoes an important boosting, mainly when the training size is insufficient. Note that this phenomenon can be found when around 4 real pages are used, and taking into account all of the above, the results are, therefore, within the expected range.

Table 2 shows the average results for all corpora and for each object-detection model for analysis purposes. The results show that the models without augmentation are not feasible in those cases in which few training pages are available, since poor figures are obtained. This table also shows that SAE achieves the best value of mAP when there are 16 or more real pages available for training, with a maximum mAP of 34.2% and 37.8%, when our data augmentation is used or not, respectively. Note that the latter occurs when using 64 real pages for training—the case with the largest number of pages within those considered. When a limited number of pages are available – 8 or fewer – Faster R-CNN seems the best option since it outperforms the performance of the rest of models. An interesting point when comparing the cases with and without image augmentation is that, on average, all the models obtain more stable figures in all cases. This is particularly interesting because it justifies the need for an algorithm with which to obtain more robust models, and this improvement is especially noteworthy when few pages are available to train them.

In either case, this experiment makes it possible to conclude that, on average, the model that should be used for LA, at least according to

Table 2
Results obtained for object detection in terms of COCO mAP (%) for scenarios with a different availability of ground-truth data. The figures in bold type indicate the best results obtained for each scenario according to the number of real pages available.

Model	Available real pages						
	1	2	4	8	16	32	64
RetinaNet							
Non-augmented	7.4	6.6	8.0	11.3	12.2	22.4	29.6
Our augmentation	23.6	24.6	25.6	25.6	26.8	26.7	25.2
Faster R-CNN							
Non-augmented	1.6	0.8	0.8	0.1	3.8	14.8	26.3
Our augmentation	**29.9**	**29.5**	**31.0**	**30.6**	30.4	30.6	31.8
SSD							
Non-augmented	1.8	1.5	1.2	0.7	4.1	3.9	10.5
Our augmentation	10.8	12.5	12.6	12.6	15.9	14.0	17.6
SAE							
Non-augmented	1.2	2.1	13.6	0.0	19.8	35.3	**37.8**
Our augmentation	23.8	26.0	21.4	27.3	**34.2**	**34.1**	31.9

Table 3
Average results in terms of mP, mR and mF_1 (%). Figures in bold type represent the best values for each metric and for each scenario considered, i.e., for a different number of available real pages. Underlined values indicate the best results for each metric, considering all the cases. Note that, because of the nature of the metrics shown in this table, the IoU threshold is necessary to determine when a predicted region can be considered as TP (see Eqs. (1) and (2). Note also that this threshold optimizes the results in the corresponding validation sets.

Scenarios	With augmentation					
	No			Yes		
	mP	mR	mF_1	mP	mR	mF_1
1 page						
RetinaNet	7.2	14.2	9.6	61.9	**62.6**	62.3
Faster R-CNN	5.2	31.0	8.9	71.1	60.5	**65.3**
SSD	5.3	20.5	8.4	56.8	46.6	51.2
SAE	**100**	2.9	5.7	62.1	28.0	38.6
2 pages						
RetinaNet	6.4	12.8	8.5	53.7	**70.6**	61.0
Faster R-CNN	45.2	25.1	32.3	**80.9**	51.6	**63.0**
SSD	71.4	17.6	28.2	53.2	46.0	49.3
SAE	18.1	6.6	9.7	66.4	38.2	48.5
4 pages						
RetinaNet	6.2	32.8	10.4	**76.5**	55.2	**64.1**
Faster R-CNN	30.2	36.0	32.8	54.8	**69.6**	61.3
SSD	61.3	11.8	19.8	59.9	38.7	47.7
SAE	36.4	23.3	28.4	38.2	32.8	35.3
8 pages						
RetinaNet	32.2	40.0	35.7	61.8	**71.6**	**66.4**
Faster R-CNN	6.7	7.9	7.2	**66.6**	66.0	66.3
SSD	31.2	6.1	10.2	64.9	23.4	34.4
SAE	0.0	0.0	0.0	40.9	36.1	38.3
16 pages						
RetinaNet	**67.9**	41.9	51.8	60.2	**77.6**	**67.8**
Faster R-CNN	51.1	31.1	38.7	65.0	69.6	67.2
SSD	52.3	28.8	37.2	42.9	42.2	42.6
SAE	42.5	31.1	35.9	50.7	55.0	52.8
32 pages						
RetinaNet	66.1	58.6	62.1	56.0	**73.4**	63.6
Faster R-CNN	43.4	49.4	46.2	**68.9**	63.2	**65.9**
SSD	67.9	9.4	16.5	45.5	45.5	45.5
SAE	53.5	54.8	54.1	54.6	51.7	53.1
64 pages						
RetinaNet	**93.4**	54.8	69.1	75.7	**68.0**	**71.6**
Faster R-CNN	61.1	64.6	62.8	75.1	57.7	65.2
SSD	75.0	29.6	42.5	65.9	55.1	60.0
SAE	53.6	57.9	55.7	51.8	51.0	51.4

mAP, depends on the training size, being Faster R-CNN the best model when up to 8 pages are available for training and SAE from there. The suitability of this metric as regards representing the retrieval of regions is analyzed in the evaluation scenario shown below, since it is the most important factor for a successful LA.

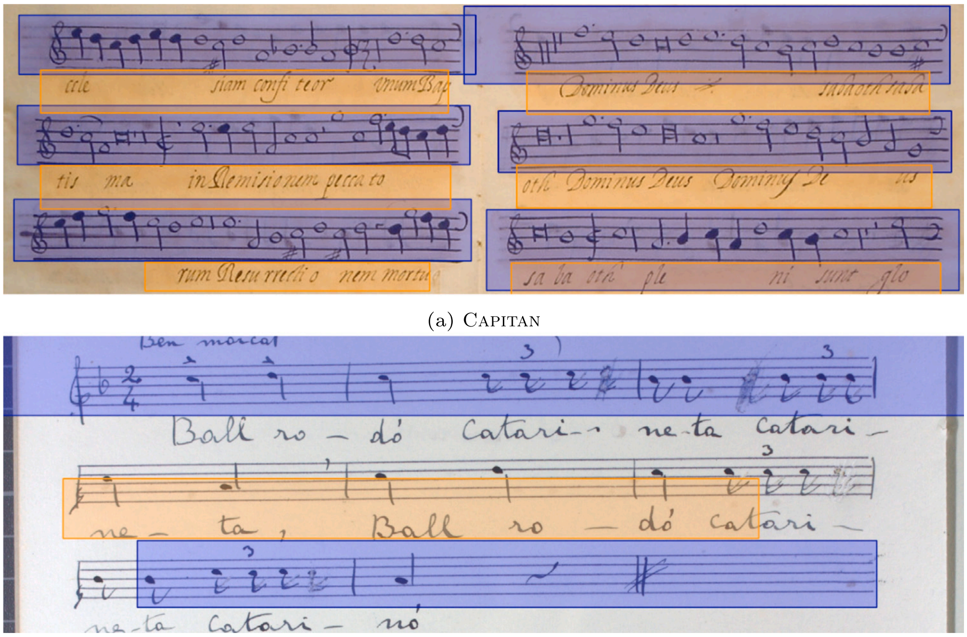
5.2. Evaluation scenario II: Evaluation in terms of regions retrieved

The mAP metric does not directly measure object detection performance. For the OMR pipeline, the detection or non-detection of regions is presumably a more relevant indicator of how useful the LA component is, rather than the extent to which the detected regions obtain high confidence values or perfectly match the ground truth region boundaries. In this scenario, we discuss the appropriateness of the popular metric mAP when used in object detection and explore whether it correlates with a greater number of detected regions, which is really the main goal of region-based LA. As mentioned above, this can be measured using mP, mR and mF_1, and we, therefore, compare the conclusions extracted from the previous evaluation scenario with those obtained by means of these metrics.

However, in order to define what a correctly detected region is, it is necessary to apply two thresholds: one for confidence and another for IoU. Confidence is, as mentioned in Section 3.1, the level of certainty that the object-detection model has when predicting the bounding boxes, whereas IoU indicates the degree of overlapping between the predicted and real regions. The results obtained for a region considered to be a correct prediction should, therefore, have sufficient confidence and IoU, signifying that those predicted regions that do not surpass these thresholds are discarded. For this reason, and because of their importance, exhaustive experimentation has been performed to obtain the best combination in the validation set. For the confidence threshold, we considered a range of values of between 0.05 and 0.95 with intervals of 0.1, while in the case of IoU, we explored common values used in object detection, specifically between 0.5 and 0.95 with a granularity of 0.05. Note that SAE does not provide any confidence value. That is, the results of this model cannot be filtered through any threshold, unlike the other models. Precisely, this is disadvantage of the SAE in practice.

We subsequently considered mP, mR and mF_1 in order to evaluate the results. Table 3 shows the average results obtained for the best combination of thresholds in each case after optimizing F_1. Table B.7 shows the specific confidence values employed to optimize that metric. These values are calculated by using all the datasets in order to obtain more generalizable conclusions. Note that confidence is a value that could be used in practice, since it is provided by the object detection model, but that the IoU can be used only in controlled scenarios, since it is computed by using the ground-truth data. These scenarios include the experiments addressed in this paper, along with the training process, since a validation set of annotated images could be used as a criterion to optimize the metrics required. The combination of both thresholds, therefore, provides a reference of the best results that could be obtained.

It will first be noted that mAP and mF_1 do not match in the model containing the best figures. As shown in Table 2, in the case of mAP, the highest value was obtained by SAE, with an average of 37.8% for 64 original pages and non-augmentation. This model also achieved the best performance with 16 or more real pages as training set, while Faster R-CNN was the best option when 8 or less training pages are available. However, as depicted in Table 3, the mF_1 metric indicates that the best results are provided by RetinaNet in combination with our data-augmentation algorithm when 64 real pages are available, with a mF_1 of 71.6%. Moreover, most of the augmented scenarios seem to indicate that RetinaNet is the best option in terms of mF_1, which is a different model to those suggested by the mAP. What is more, the mF_1 figures are considerably higher than the mAP figures, and, as will be



(a) Capitan

(b) FMT-C

Fig. 4. Selected extractions of object detection by using RetinaNet, in which correct and incorrect estimations are shown. Blue bounding boxes represent predictions of staff regions, whereas the orange boxes represent the lyric areas obtained.

seen later in Fig. 4, visually, higher values are more correlated with the regions retrieved.

Focusing on the results, it will be observed that the augmentation algorithm proposed is crucial in terms of mF_1. In none of the non-augmented cases does this metric supersede the augmented scenarios, and this, therefore, justifies the theory that this algorithm is able to increase the robustness of the models as regards extracting the bounding boxes. There are some examples of the non-augmented experiment in which mP improves the results obtained after augmentation, particularly in the case of SAE with 1 page, which obtains 100%–because of the lack of FP in this case–while RetinaNet yields 67.9% and 93.4% for 16 and 64 pages, respectively. However, when focusing on the values of mR for these cases, it will be noted that SAE attains only 2.9%, and RetinaNet obtains 41.9% and 54.8%, respectively. This signifies that, in these cases, the models prioritize the detection of real regions over the miss-detection of regions in which there is no information. In other words, there are regions that have to be manually discarded. This situation could be interesting depending on the task being carried out, but a balanced model could, in general terms, be beneficial as regards obtaining good results with less human intervention.

As shown in Table 3, the augmented scenarios attain more stable and balanced results in terms of mP and mR, impacting directly on better mF_1 figures. Indeed, all the experiments evaluated using this metric with augmentation were, on average, better than the non-augmented cases. According to these metrics, the best model for LA is RetinaNet being a potential solution since it obtains the highest mF_1 value – 71.6% – with 64 pages, and the best mR with a value of 77.6%.

However, in the experiment shown in Section 5.1, Faster R-CNN and SAE were reported as the best models; therefore, the conclusions from the two experiments are different. It is precisely this situation which justifies our claim that mAP metric might not be suitable for evaluating LA, since the real importance of this process is the extraction of bounding boxes as discrete objects, and not only the evaluation of their overlapping and confidence values.

In order to complement the LA experiments, Fig. 4 shows an example from Capitan in which the bounding boxes are correctly predicted and another example from FMT-M in which there are several miss-detections.

In the first example, which is shown in Fig. 4(a), it will be observed that the retrieval region is generally of good quality. In this example, the bounding boxes retrieved appear to correctly contain the relevant information, but, it should be noted that the staff retrieval in this example obtains an average IoU of 79% and a confidence of 55%, whereas the text retrieval obtains an IoU of 74% and a confidence of 39%. This signifies that, although the prediction of the bounding would appear to be graphically suitable and correct, since the objects are large and regular, a slight error, especially on the vertical side in our context, may considerably worsen the IoU. This, therefore, means that it is not necessary to attain a perfect matching of IoU in order to cover the data that has to be retrieved, and this also explains why the mAP obtains significantly lower figures, since the range between 80% to 95% would not, on average, contain any bounding boxes. Moreover, the confidence of the model is, on average, very poor when compared to what might be expected after visual inspection, signifying that a particularly low confidence threshold is needed for this metric in order to prevent these regions from being discarded.

In the second example depicted in Fig. 4(b), there are certain issues as regards both staff and text retrieval. In visual terms, the staff at the top appears to have been correctly retrieved, although the area detected is higher than the staff itself. The principal problems with the staves in this case are that one staff is not detected, the last one is partially retrieved, and two music symbols are missed. These issues are crucial for the eventual transcription, and a manual correction would, therefore, be required in order to correctly extract the bounding boxes. With regard to the text regions, only one bounding box is retrieved, but it does not cover the text at all. Two other text lines are not detected, and manual corrections would, therefore, also have to be performed for a full digitization. In this example, the staff predictions have 57% of average IoU with respect to the ground truth, whereas RetinaNet provides an average confidence of only 28%. In the case of the text regions, the IoU obtained is 25%, despite the fact that the text is quite well detected, and the confidence reaches 59%.

This qualitative analysis, therefore, reinforces the idea that it is not necessary to obtain a perfect matching of the bounding boxes, and that it is sufficient to obtain regions that encompass the content. This demonstrates that mAP is not an appropriate metric with which to evaluate the objects retrieved, since it places much more importance on the overlapping with the ground truth. It is consequently possible to conclude that the mF_1 metric is more suitable than mAP for the LA of music score images, despite the popularity of mAP in object detection. To complete and confirm our analysis, in the next evaluation scenario, we further analyze the influence of IoU and confidence on the final transcription.

5.3. Evaluation scenario III: Goal-directed evaluation

We have, until this point, performed a thorough analysis of the LA stage on its own, without any specific context. However, it is important to recall that this stage is not, in most cases, an objective in itself, but merely an intermediate step within a pipeline employed to transcribe the content of music score images. In this section we, therefore, study the relationship between the operation of LA and the transcription process itself, focusing particularly on the music notation (regions with staves). To this end, we selected RetinaNet as being representative of an automatic layout analysis stage, given that it was, according to our previous experiment, the best option.

For this goal-directed experiment, we employed a state-of-the-art model for OMR that was built as a Convolutional Recurrent Neural Network (CRNN) and was directly trained to retrieve the sequence of musical symbols found in the image of a single staff. Since the CRNN is used here as a black box, the reader is referred to a number of works for further details on its operation (Calvo-Zaragoza et al., 2019; Shi et al., 2016; Wick & Puppe, 2021).

The experiment outlined in this section is as follows:

1. For each corpus, and using the training and validation partitions, the CRNN is trained by means of the ground-truth regions along with their corresponding transcripts, thus ensuring the best possible recognition model.
2. With regard to the test partitions, we employ the LA model to automatically retrieve the staff regions, along with their confidence.
3. Each predicted staff is matched with all the ground-truth regions of the test partition for which the IoU is greater than 0.55. This specific value was the one that maximized the mF_1 over the validation set in the previous experiment.
4. For each match, both the detected and the ground-truth staves are processed with the CRNN in order to retrieve their music symbols.

We denote as $\overline{\text{SER}}$ the difference in SER between the symbols retrieved from the ground-truth staff and the symbols retrieved from the detected staff. This will be used as a measure of the impact of the layout analysis: if 0, this signifies that there is no actual difference between retrieving the content using the manually-annotated region and retrieving the content using the automatically-detected region (a fairly ideal scenario). As this difference grows, the loss in performance caused by the layout analysis is greater. In turn, it might occur that the SER is smaller in the region predicted automatically, signifying that the $\overline{\text{SER}}$ would be negative. Whatever the case may be, for each detected region, we eventually obtain a tuple ($\overline{\text{SER}}$, confidence, IoU).

Furthermore, before reporting the results of this experiment, it should be taken into account that some deviations in the detected regions could be alleviated by training the CRNN with data augmentation by, for example, slightly modifying the corners of the training staff regions. The effect of data augmentation on staff-based OMR with CRNN has already been studied in previous works (López-Gutiérrez et al., 2021), although not comprehensively in the context of its connection with an imperfect layout analysis. Here we shall, therefore, consider the CRNN with and without this type of data augmentation in order to also carry out the study from this perspective.

Fig. 5 shows the contrast of the confidence and IoU values (x-axes) of the detected regions with the $\overline{\text{SER}}$ (y-axes), highlighting the different corpora. An initial remark is that, as might be expected, the LA has less impact on the regions that have a higher confidence and a higher IoU (right-hand side of the images), in which the $\overline{\text{SER}}$ is closer to 0. As the model has less confidence in the regions or the IoU decreases, this value clearly increases. This even produces cases of $\overline{\text{SER}} = 1$, signifying that the CRNN perfectly retrieves the symbols for the ground-truth region but completely fails in the case of the predicted one.

Table 4
Pearson's correlation coefficient between IoU and $\overline{\text{SER}}$ and between the confidence value and $\overline{\text{SER}}$. The columns "No" and "Yes" indicate the use of the data augmentation mechanism.

Corpus	IoU - $\overline{\text{SER}}$		Confidence - $\overline{\text{SER}}$	
	No	Yes	No	Yes
Capitan	−0.87	−0.67	−0.82	−0.65
Seils	−0.81	−0.67	−0.88	−0.66
FMT-M	−0.40	−0.59	−0.06	−0.40
FMT-C	−0.11	−0.34	−0.07	−0.28

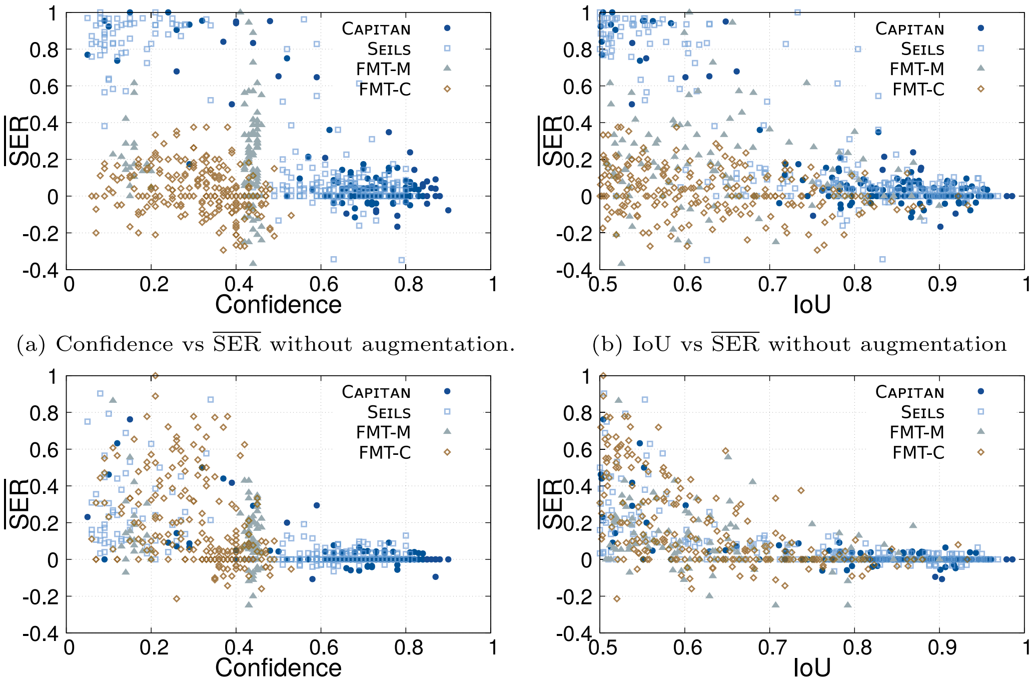
The aforementioned phenomenon has a double reading, particularly in the case of confidence (Figs. 5(a) and 5(c)): while it is true that the results are quite poor when transcribing the staff from the less reliable regions, these could easily be discarded. The full OMR system should consider only those regions for which the confidence is high and for which a positive result is, in most cases, expected. For this LA model, and for all corpora in general, it appears that a suitable threshold for such a purpose would be 0.6.

Furthermore, the correlation between the IoU and the $\overline{\text{SER}}$ also produces a clear trend (Figs. 5(b) and 5(d)): the higher the IoU, the lower the deterioration of the transcription. Unlike confidence, this case cannot be predicted in practice, since the IoU can be computed only in controlled experiments in which the true bounding box of a region is known. However, these results could serve to better validate the models in training time. In this case, the threshold beyond which the results drastically change the $\overline{\text{SER}}$ depends on the CRNN that is used, as discussed below.

If we compare the results of the base CRNN (Figs. 5(a) and 5(b)) with a CRNN trained with data augmentation for the regions (Figs. 5(c) and 5(d)), it is clear that the latter is more robust to an (imperfect) automatic LA, as would be expected with this type of techniques. In the case of confidence, there is not much difference; however, in the case of IoU, the results are notably better. While without data augmentation, the threshold for which the results are reliable is around 0.9 – signifying that an almost perfect match is required – the data augmentation manages to enable the CRNN to correctly recover the musical symbols with an IoU of above approximately 0.7, signifying a much more reasonable case to attain in practice.

It should be noted that FMT-C has a different trend for the lowest confidence values. This phenomenon is aligned with the fact that data augmentation improves the results for ground-truth staves, but the predicted regions do not undergo this improvement. However, as seen in Section 5.1 (and also supported by Appendix B), the corpus with the worst quality in the layout analysis process is FMT-C. This means that the predicted regions do not have the quality required to obtain a usable model. We attribute this low quality to the density and degree of overlapping of consecutive regions, thus affecting the region-retrieval process and, in turn, making it more difficult to recognize the sequence.

Finally, we also provide a correlation measure in order to verify the previous analysis. Table 4 shows the Pearson's correlation between the variables that appear in the graphs in Fig. 5, which is a typical metric for measuring the relationship between two variables. These correlations are bounded between −1 and 1; values closer to 1 represent high correlation, values closer to 0 denote lack of correlation, and values closer to −1 indicate inverse correlation. As observed in the table, all the figures are negative, indicating an inverse correlation. This is to be expected, because $\overline{\text{SER}}$ represents the error made by the end-to-end approach, and, as seen in Fig. 5, this error generally increases



(a) Confidence vs $\overline{\text{SER}}$ without augmentation.

(b) IoU vs $\overline{\text{SER}}$ without augmentation

(c) Confidence vs $\overline{\text{SER}}$ with augmentation.

(d) IoU vs $\overline{\text{SER}}$ with augmentation

Fig. 5. Relationship between IoU and confidence obtained for LA and the music transcription, in this case scored using $\overline{\text{SER}}$.

when IoU and confidence decrease, leading to an inverse relationship between these variables. This causes the correlation coefficient to be negative in all cases. In addition, note that Capitan and Seils attain particularly high negative correlations for both measures, without and with data augmentation.

Focusing on the case without data augmentation, the correlations between IoU and $\overline{\text{SER}}$ are −0.87 and −0.81 for Capitan and Seils, whereas the correlations between confidence and $\overline{\text{SER}}$ are −0.82 and −0.88. These are the cases where the lowest values are achieved, which corroborate the fact that these two datasets obtain better results in region retrieval, and therefore, the quality of the regions is usually high. As the regions are properly recovered with high IoU and confidence, the end-to-end approach is able to obtain low SER values for both the predicted staves and the ground-truth staves, signifying that $\overline{\text{SER}}$ is also low. This explains the correlation values obtained in these cases. In addition, when data augmentation is applied, the IoU-$\overline{\text{SER}}$ correlations achieve −0.67 for both datasets, and the confidence-$\overline{\text{SER}}$ correlations obtain −0.65 for Capitan and −0.66 for Seils. Despite this increase in correlation, the values continue to be negative and, as shown in Fig. 5, data augmentation improves the results for Capitan and Seils. This means that, as expected, the IoU and confidence of the regions lose importance when data augmentation is applied, since, although the correlations are slightly higher than in the non-augmented cases, the $\overline{\text{SER}}$ results are better.

There are also negative correlations in the case of the FMT datasets, although not as extreme as in the cases explained above. This supports that as the quality of the regions retrieved in these cases are significantly worst in terms of both IoU and confidence, the SER values are not so low. There is, therefore, less difference between the variables, and this makes the correlation values higher than in the case of Capitan and Seils. In the non-augmented case, there are IoU-$\overline{\text{SER}}$ correlations of −0.4 and −0.11 for FMT-M and FMT-C, respectively, whereas the confidence-$\overline{\text{SER}}$ correlations are −0.06 and −0.07. In these cases, when data augmentation is applied, these values decrease to −0.59 and −0.34 for the IoU-$\overline{\text{SER}}$ correlations, and −0.4 and −0.28 for the confidence-$\overline{\text{SER}}$ correlations. Because of the difficulty of using these datasets in the two processes studied (region retrieval and end-to-end recognition), the SER metrics are significantly worse (higher) when compared with Capitan and Seils, even for ground-truth staves. In these cases, as the SER values are high, they can easily be improved, and indeed, the results of the end-to-end often improve when data augmentation is applied, signifying that both predicted and ground-truth staves undergo a boosting in this music-sequence recognition, and in these cases, the ground-truth results improve to a greater extent. Note that the IoU and confidence of the predicted staves are extremely low, meaning that even if data augmentation is applied to the end-to-end model, a limited improvement can be observed in predicted staves. This improvement is significantly greater in the case of the ground-truth staves, and $\overline{\text{SER}}$, therefore, generally increases their values. As the IoU and confidence values are low, and $\overline{\text{SER}}$ increases, the correlation values decrease in these cases.

In conclusion, when the staff retrieval has high quality, the end-to-end model attains high-performance results as regards music-notation recognition, even if no data augmentation is applied. However, when data augmentation is applied, the correlation between IoU/confidence and $\overline{\text{SER}}$ increases slightly, signifying that when sufficient quality in the staff recognition is achieved, the perfection with which they have been detected is less important. This occurs in the cases of Capitan and Seils. However, when a difficult corpus such as FMT-M or FMT-C is employed, the recognition of the staves is significantly weaker, and the recognition of the music sequence is, therefore, affected. Even if data augmentation is applied to the end-to-end model, the predicted staves do not undergo a relevant improvement, although the ground-truth staves may undergo important boosting. This coincides with the analysis of Fig. 5, since low IoU and confidence values often coincide with high $\overline{\text{SER}}$ values.

6. Conclusions

This work presents comprehensive experiments carried out in order to assess the region-based LA process for music score images. This was done by carrying out three specific evaluation scenarios in which different aspects and goals were assessed.

The first scenario focused on an analysis of the behavior of several well-known object-detection models in different scenarios according to the availability of ground-truth data, which are often scarce. In order to palliate this situation when few annotated images are provided, we have proposed and evaluated a data-augmentation algorithm with which to generate semi-synthetic images from the bounding boxes of the original pages. In this scenario, we considered a common metric used in object detection in multiple contexts: the mAP. When detecting the different regions, the results obtained when employing this metric suggest that Faster R-CNN is the best option when 8 or less real pages are available for training the model, while the SAE is the best option when the training set has at least 16 real pages. In addition, of all models considered, SAE provided the highest mAP value. RetinaNet also obtains competitive mAP figures, whereas the model with the lowest performance when this metric is employed is the SSD.

The objective of the second scenario was to demonstrate that the metric considered previously – mAP – is not necessarily the best means of evaluating models for the LA of music score images. This metric addresses the assessment as an overlapping problem; however, in OMR, the number of predicted regions considered as being correct is even more crucial than the overlap between the predicted and the real bounding boxes, as long as the relevant information is included within these regions. We have, therefore, carried out an evaluation by using the macro average versions of precision, recall and f-score—mP, mR and mF_1, respectively. After the analysis, the model that obtained the highest mR and mF_1 was RetinaNet, with 77.6% and 71.6%, respectively, and it generally attained more stable and balanced figures for all the models. This also proves that the results obtained by the mAP metric do not coincide with those extracted from mF_1. Whereas mAP prioritizes the matching of the area with the ground truth, mF_1 does so with the region retrieval. Note, however, that prioritizing the matching of the area depends on a subjective component, since the ground truth is often obtained by hand, and, as studied in previous work (Castellanos et al., 2020), a staff-level music-sequence recognition trained with "perfect" bounding boxes may not be aligned with better transcriptions since the predicted regions will not be perfect either. It is, therefore, possible to conclude that F_1-based metrics are more aligned with that which is required in OMR systems.

With regard to the third scenario, we attempted to evaluate the relationship between the overlapping of predicted and annotated staff regions, the confidence provided by the LA model, and the error obtained in the final transcription through the use of an end-to-end strategy by means of CRNN, measured with the SER metric. We additionally explored the influence of the data augmentation shown in previous works on these relationships. One of the main conclusions obtained was that high confidence and IoU values are strongly aligned with low transcription errors. Indeed, in the case of all the corpora evaluated, we observed an abrupt reduction in the error from a certain value of confidence and IoU. This supports the idea that using thresholds to filter the LA regions is a correct way in which to discard those regions that may cause errors in the transcription. This ensures a certain quality of the transcriptions, which could be used to train other end-to-end models.

As the aforementioned results show, no model detects all the regions of interest in music score images. A specific object-detection model for LA in OMR could be a promising avenue for further research, in which specific characteristics of this type of documents could be exploited, such as the fact that their general structure is regular or that their regions are usually wider than taller. Another interesting aspect of this topic that could be explored is the usage of transfer learning techniques, along with an analysis of the contribution of pre-trained models in OMR, since there is a gap in literature in this context. A generic music corpus with which to perform this task could also be created. Furthermore, it would be interesting to evaluate the performance of these models in cross-manuscript cases (a model trained for one collection and used in another) and to propose improvement strategies in this regard using un- or semi-supervised domain adaptation techniques.

CRediT authorship contribution statement

Francisco J. Castellanos: Conceptualization, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Carlos Garrido-Munoz:** Methodology, Software, Validation, Writing-reviewing. **Antonio Ríos-Vila:** Methodology, Software, Validation, Writing-reviewing. **Jorge Calvo-Zaragoza:** Conceptualization, Investigation, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This paper is part of the I+D+i PID2020-118447RA-I00 (MultiScore) project funded by MCIN/AEI/10.13039/501100011033, Spain and the GV/2020/030, Spain project funded by the Generalitat Valenciana, Spain. The first and third authors acknowledge support from the "Programa I+D+i de la Generalitat Valenciana, Spain" through grants ACIF/2019/042 and ACIF/2021/356, respectively.

Appendix A. Configuration of the neural networks

This appendix shows the configuration of the neural network models used in our experiments.

A.1. RetinaNet

We have considered the PyTorch implementation[3] of RetinaNet pre-trained on COCO 2017. The number of output classes of the model was changed in order to retrieve staves and lyrics. We used the same hyper-parameters as in the original paper (Lin et al., 2020). The backbone used was a 50-layer Residual Net (He et al., 2016) with its original hyper-parameters. Input images were resized internally to 224 × 224 px.

A.2. Faster R-CNN

We used the PyTorch implementation[4] of Faster-RCNN with Feature Pyramids (Ren et al., 2015), pre-trained on COCO 2017. The number of output classes of the model was changed to match our requirements (staves and lyrics retrieval). We used the same hyper-parameters shown in Ren et al. (2015). The anchors used were 3 scales with box areas of 128, 256 and 512 squared pixels and 1:1, 1:2 and 2:1 for RPN (Region Proposal Network), as in the original paper. The backbone used was a 50-layer Residual Net (He et al., 2016), and its original hyper-parameters were maintained. As in the case of RetinaNet, the images were resized internally to 224 × 224 px.

A.3. SSD

With regard to the architecture used for SSD, we employed the PyTorch implementation,[5] details of which can be found in the original paper (Liu et al., 2016), and pre-trained on COCO 2017. The number of output classes of the model was modified in order to retrieve staves and lyrics. The backbone used was a VGG-16 (Simonyan & Zisserman, 2015). The images were internally resized to 224 × 224 px.

[3] http://torchvision.models.detection.retinanet_resnet50_fpn.

[4] http://torchvision.models.detection.fasterrcnn_resnet50_fpn.

[5] http://torchvision.models.detection.ssd300_vgg16.

Table B.5
Quantity of staves (%) that are considered as TP. The columns "No" and "Yes" indicate the use of our data augmentation mechanism.

Corpus	Available real pages without (No) and with (Yes) data augmentation													
model	**1**		**2**		**4**		**8**		**16**		**32**		**64**	
	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Capitan														
RetinaNet	83.3	87.5	70.8	87.5	62.5	90.3	83.3	84.7	86.8	90.3	88.9	88.9	99.3	83.3
Faster R-CNN	0.0	93.1	49.3	99.3	1.4	99.3	0.0	99.3	81.9	100	99.3	100	99.3	100
SSD	42.4	83.3	42.4	77.1	43.8	86.8	23.6	82.6	60.4	84.7	63.9	86.8	79.2	85.4
SAE	0.0	75.1	0.0	63.4	66.0	77.4	0.0	72.8	19.0	75.6	38.2	71.2	20.8	81.5
Seils														
RetinaNet	0.0	86.3	0.0	85.2	61.2	83.5	75.9	82.5	84.5	86.3	87.3	80.8	89.0	83.5
Faster R-CNN	51.5	85.9	0.0	88.7	51.9	85.2	0.0	89.3	0.0	83.2	69.8	83.2	84.5	88.3
SSD	52.6	81.8	50.5	83.2	47.8	80.4	0.0	81.4	59.5	82.5	0.0	84.2	84.5	80.8
SAE	0.0	83.8	0.0	88.3	21.0	0.0	0.0	71.1	89.7	84.5	89.7	64.9	90.0	76.6
FMT-M														
RetinaNet	0.0	61.6	0.0	66.9	70.1	68.5	59.3	55.3	0.0	63.8	71.7	69.0	31.2	48.7
Faster R-CNN	2.4	65.1	0.0	85.4	0.0	86.5	23.0	81.5	88.9	84.1	73.3	83.6	75.9	78.3
SSD	0.0	56.6	0.0	59.3	0.0	74.9	0.0	0.0	52.4	68.5	7.7	74.6	0.0	66.9
SAE	0.0	19.3	35.4	20.4	75.1	60.6	0.0	45.2	51.9	80.7	58.5	83.1	62.2	77.0
FMT-C														
RetinaNet	49.2	81.7	60.5	74.0	69.5	85.5	73.6	88.7	71.1	86.8	71.7	83.0	80.1	83.6
Faster R-CNN	63.7	83.0	70.1	89.4	70.7	85.5	0.0	86.8	81.0	80.4	78.1	85.2	79.1	83.6
SSD	48.9	0.0	36.7	0.0	0.0	0.0	16.7	0.0	29.6	0.0	0.0	0.0	0.0	55.0
SAE	19.0	0.0	0.0	17.4	0.0	22.8	0.0	19.3	28.6	23.2	26.7	19.9	27.7	21.2

Table B.6
Quantity of lyrics (%) that are considered as TP. The columns "No" and "Yes" indicate the use of our data augmentation mechanism.

Corpus	Available real pages without (No) and with (Yes) data augmentation													
model	**1**		**2**		**4**		**8**		**16**		**32**		**64**	
	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Capitan														
RetinaNet	1.0	92.2	1.2	96.6	2.1	94.8	82.8	96.6	82.8	96.6	87.1	95.7	88.8	87.9
Faster R-CNN	21.6	95.7	0.9	94.8	0.9	93.1	0.0	95.7	0.0	95.7	80.2	96.6	80.2	95.7
SSD	36.2	70.7	40.5	64.7	31.0	79.3	47.4	77.6	47.4	77.6	50.9	69.0	66.4	84.5
SAE	0.0	91.2	0.0	92.9	0.0	88.2	0.0	93.9	0.0	90.7	40.2	91.7	19.8	83.8
Seils														
RetinaNet	0.0	59.6	0.0	66.1	0.0	69.3	31.2	61.9	31.2	61.9	85.7	68.8	74.9	70.9
Faster R-CNN	0.0	41.5	1.5	57.7	0.0	47.8	1.1	43.4	1.1	43.4	33.9	55.6	66.9	61.3
SSD	0.0	41.1	0.0	52.8	0.0	55.4	0.0	33.9	0.0	33.9	0.0	55.2	57.9	55.6
SAE	0.0	5.5	0.0	30.9	0.0	30.5	0.0	0.0	0.0	46.1	46.9	30.5	49.1	37.9
FMT-M														
RetinaNet	0.0	40.8	0.0	37.6	0.0	55.5	0.0	65.3	0.0	65.3	57.6	62.4	71.4	35.5
Faster R-CNN	0.0	39.6	0.4	12.7	0.4	27.8	0.4	46.1	0.4	46.1	17.6	38.4	32.7	33.1
SSD	0.0	34.3	0.0	19.2	0.0	8.6	0.0	0.0	0.0	0.0	0.0	28.6	0.0	38.8
SAE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	32.2	0.0	15.5	46.1	42.4	59.2	26.5
FMT-C														
RetinaNet	0.0	54.8	0.0	14.0	0.0	22.6	0.0	22.6	0.0	22.6	0.0	17.2	9.7	23.7
Faster R-CNN	0.0	50.5	0.0	55.9	0.0	31.2	4.3	49.5	4.3	49.5	0.0	55.9	5.4	23.7
SSD	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	30.1
SAE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table B.7
Confidence threshold that optimizes the F_1 metric in the validation set. The columns "No" and "Yes" indicate the use of our data augmentation mechanism. Note that the SAE model does not provide confidence values in its output, and this model is not, therefore, included in this table.

Type of region	Available real pages without (No) and with (Yes) data augmentation													
Model	**1**		**2**		**4**		**8**		**16**		**32**		**64**	
	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Staves														
RetinaNet	0.25	0.35	0.25	0.35	0.15	0.35	0.25	0.35	0.25	0.35	0.25	0.35	0.35	0.35
Faster R-CNN	0.05	0.65	0.05	0.75	0.05	0.65	0.45	0.65	0.15	0.75	0.15	0.55	0.35	0.85
SSD	0.15	0.15	0.15	0.15	0.15	0.25	0.15	0.15	0.15	0.25	0.15	0.15	0.15	0.15
Lyrics														
RetinaNet	0.05	0.35	0.05	0.25	0.05	0.35	0.15	0.25	0.15	0.25	0.15	0.25	0.25	0.25
Faster R-CNN	0.35	0.35	0.45	0.35	0.45	0.15	0.05	0.25	0.15	0.25	0.15	0.25	0.15	0.25
SSD	0.05	0.05	0.15	0.05	0.15	0.15	0.35	0.15	0.15	0.15	0.15	0.15	0.15	0.15

A.4. SAE

The implementation considered for the SAE architecture is defined in Castellanos et al. (2020). The model consists of a U-net architecture with encoder and decoder sections. The encoder part includes 3 blocks in which each block has a convolution of 128 filters and a kernel filter of 5×5 of size, a Rectifier Linear Unit (ReLU) activation and a max-pooling operator, and the decoder part also contains 3 blocks in which each block contains a convolution layer with the same properties as that used in the encoder: a ReLU activation and an up-sampling operator. Finally, a convolutional layer of one filter and the same kernel and a sigmoid activation was applied in order to provide the probabilistic map in which each element contains the probability of the pixel being part of a region. As occurred in the paper mentioned, the input images were resized to 512×512 px. before being processed by the neural network. After this processing, the result was then resized again to match with the original image. Finally, a connected component analysis was performed to retrieve the coordinates of the bounding boxes with respect to the original image.

Note that, because this type of architecture limits its predictions to one class per pixel, it may experience difficulties in cases with a high degree of overlapping between the bounding boxes. In order to lessen this issue, we reduced the vertical size of the bounding boxes by 20% before the first resizing of the image, as proposed in Castellanos et al. (2020). The size was then recovered after the detection of the coordinates of the predicted bounding boxes.

A.5. Convolutional recurrent neural network

In the last experiment, we make use of a Convolutional Recurrent Neural Network (CRNN) to retrieve the transcription (sequence of music symbols) of each detected staff. Its configuration follows the best hyper-parameterization found in the work of Calvo-Zaragoza et al. (2019): four convolutional, two recurrent layers, and one dense layer, trained by employing the Connectionist Temporal Classification loss function. The hyper-parameters of each layer are detailed in the original paper.

Appendix B. Supplementary layout analysis results

The results provided in this appendix complement those shown in Section 5. First, Tables B.5 and B.6 provide a reference to the amount of regions that have been recognized, measured in % for both staves and lyrics, respectively. These tables, therefore, represent the number of regions that can be interpreted as TP in the computation of the metrics used in Section 5.2, i.e., precision, recall and the F-score metric.

Table B.7 shows the confidence threshold used for the computation of the metrics used in Section 5.2. These values are those that optimize the F-score value in the validation set for all the corpora considered. This table provides the thresholds used for the cases with and without data augmentation, according to the number of real pages available in the training set.

References

Binmakhashen, G. M., & Mahmoud, S. A. (2019). Document layout analysis: A comprehensive survey. *ACM Computing Surveys*, *52*(6), 1–36.

Bosch, V., Calvo-Zaragoza, J., Toselli, A. H., & Vidal-Ruiz, E. (2016). Sheet music statistical layout analysis. In *15th International conference on frontiers in handwriting recognition* (pp. 313–8).

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010* (pp. 177–186). Springer.

Burgoyne, J. A., Ouyang, Y., Himmelman, T., Devaney, J., Pugin, L., & Fujinaga, I. (2009). Lyric extraction and recognition on digital images of early music sources. In *Proceedings of the 10th International society for music information retrieval conference, vol. 10* (pp. 723–727).

Calvo-Zaragoza, J., Castellanos, F. J., Vigliensoni, G., & Fujinaga, I. (2018). Deep neural networks for document processing of music score images. *Applied Sciences*, *8*(5), 654.

Calvo-Zaragoza, J., Hajič, J., Jr., & Pacha, A. (2020). Understanding optical music recognition. *ACM Computing Surveys*, *53*(4).

Calvo-Zaragoza, J., Rizo, D., & Quereda, J. M. I. (2016). 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* (pp. 509–514).

Calvo-Zaragoza, J., Toselli, A. H., & Vidal, E. (2019). Handwritten music recognition for mensural notation with convolutional recurrent neural networks. *Pattern Recognition Letters*, *128*, 115–121.

Castellanos, F. J., Calvo-Zaragoza, J., & Iñesta, J. (2020). A neural approach for full-page optical music recognition of mensural documents. In *Proceedings of the 21st International society for music information retrieval conference* (pp. 558–565). Montréal, Canada.

Castellanos, F. J., Calvo-Zaragoza, J., Vigliensoni, G., & Fujinaga, I. (2018). Document analysis of music score images with selectional auto-encoders. In *Proceedings of the 19th International society for music information retrieval conference* (pp. 256–263).

Dalitz, C., Droettboom, M., Pranzas, B., & Fujinaga, I. (2008). A comparative study of staff removal algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(5), 753–766.

Géraud, T. (2014). A morphological method for music score staff removal. In *2014 IEEE International conference on image processing* (pp. 2599–2603). IEEE.

Girshick, R. (2015). Fast R-CNN. In *Proceedings of the IEEE International conference on computer vision* (pp. 1440–1448).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press, http://www.deeplearningbook.org.

He, S., & Schomaker, L. (2019). DeepOtsu: Document enhancement and binarization using iterative deep learning. *Pattern Recognition*, *91*, 379–390.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on computer vision and pattern recognition* (pp. 770–778). IEEE Computer Society.

Huang, Z., Jia, X., & Guo, Y. (2019). State-of-the-art model for music object recognition with deep learning. *Applied Sciences*, *9*(13), 2645–2665.

Jia, X., Song, Y., Ma, S., & Ding, P. (2021). Printed score detection based on deep learning. In *2021 Asia-Pacific Conference on communications technology and computer science* (pp. 173–177). IEEE.

Kletz, M., & Pacha, A. (2021). Detecting staves and measures in music scores with deep learning. In J. Calvo-Zaragoza, & A. Pacha (Eds.), *Proceedings of the 3rd International workshop on reading music systems* (pp. 8–12). Alicante, Spain.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lin, T., Goyal, P., Girshick, R. B., He, K., & Dollár, P. (2020). Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(2), 318–327.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740–755). Springer.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single shot MultiBox detector. In *European conference on computer vision* (pp. 21–37). Springer.

Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., & Pietikäinen, M. (2020). Deep learning for generic object detection: A survey. *International Journal of Computer Vision*, *128*(2), 261–318.

López-Gutiérrez, J. C., Valero-Mas, J. J., Castellanos, F. J., & Calvo-Zaragoza, J. (2021). Data augmentation for end-to-end optical music recognition. In E. H. Barney Smith, & U. Pal (Eds.), *Document analysis and recognition – ICDAR 2021 workshops* (pp. 59–73).

Neubeck, A., & Van Gool, L. (2006). Efficient non-maximum suppression. In *18th International conference on pattern recognition, vol. 3* (pp. 850–855). IEEE.

Pacha, A. (2019). Incremental supervised staff detection. In *Proceedings of the 2nd International workshop on reading music systems, delft, the Netherlands* (pp. 16–20).

Pacha, A., Hajič, J., & Calvo-Zaragoza, J. (2018). A baseline for general music object detection with deep learning. *Applied Sciences*, *8*(9), 1488.

Parada-Cabaleiro, E., Batliner, A., & Schuller, B. W. (2019). A diplomatic edition of il lauro secco: Ground truth for OMR of white mensural notation. In *ISMIR* (pp. 557–564).

Pastor-Pellicer, J., Boquera, S. E., Zamora-Martínez, F., Afzal, M. Z., & Bleda, M. J. C. (2015). Insights on the use of convolutional neural networks for document image binarization. In *Advances in computational intelligence - 13th international work-conference on artificial neural networks* (pp. 115–126).

Quirós, L., Toselli, A. H., & Vidal, E. (2019). Multi-task layout analysis of handwritten musical scores. In *Iberian conference on pattern recognition and image analysis* (pp. 123–134). Springer.

Rebelo, A., Fujinaga, I., Paszkiewicz, F., Marçal, A. R. S., Guedes, C., & Cardoso, J. S. (2012). Optical music recognition: state-of-the-art and open issues. *The International Journal of Multimedia Information Retrieval*, *1*(3), 173–190.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems, December 7-12, Montreal, Quebec, Canada* (pp. 91–99).

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Medical image computing and computer-assisted intervention – MICCAI 2015* (pp. 234–241). Cham: Springer International Publishing.

Ros-Fábregas, E. (2021). Codified spanish music heritage through verovio: the online platforms fondo de música tradicional IMF-CSIC and books of hispanic polyphony IMF-CSIC . In *Proceedings of the music encoding conference*. Alicante, Spain.

dos Santos Cardoso, J., Capela, A., Rebelo, A., Guedes, C., & da Costa, J. P. (2009). Staff detection with stable paths. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(6), 1134–1139.

Sauvola, J., & Pietikäinen, M. (2000). Adaptive document image binarization. *Pattern Recognition*, *33*(2), 225–236.

Shi, B., Bai, X., & Yao, C. (2016). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(11), 2298–2304.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Y. Bengio, & Y. LeCun (Eds.), *3rd International conference on learning representations*.

Studer, L., Alberti, M., Pondenkandath, V., Goktepe, P., Kolonko, T., Fischer, A., Liwicki, M., & Ingold, R. (2019). A comprehensive study of ImageNet pre-training for historical document image analysis. In *2019 International conference on document analysis and recognition* (pp. 720–725). http://dx.doi.org/10.1109/ICDAR.2019.00120.

Waloschek, S., Hadjakos, A., & Pacha, A. (2019). Identification and cross-document alignment of measures in music score images. In *20th International society for music information retrieval conference* (pp. 137–143).

Wick, C., & Puppe, F. (2021). Experiments and detailed error-analysis of automatic square notation transcription of medieval music manuscripts using CNN/LSTM-networks and a neume dictionary. *Journal of New Music Research*, *50*(1), 18–36.

Zhong, X., Tang, J., & Jimeno Yepes, A. (2019). PubLayNet: Largest dataset ever for document layout analysis. In *2019 International conference on document analysis and recognition* (pp. 1015–1022). http://dx.doi.org/10.1109/ICDAR.2019.00166.