Multi-Task Layout Analysis of Handwritten Musical Scores

Lorenzo Quirós[1], Alejandro H. Toselli[1], and Enrique Vidal[1]

Pattern Recognition and Human Language Technologies Research Center
Universitat Politècnica de València, Camino de Vera, s/n, 46022, Valencia, Spain
{loquidia,ahector,evidal}@prhlt.upv.es
https://www.prhlt.upv.es

Abstract. Document Layout Analysis (DLA) is a process that must be performed before attempting to recognize the content of handwritten musical scores by a modern automatic or semiautomatic system. DLA should provide the segmentation of the document image into semantically useful region types such as staff, lyrics, etc. We present a system that extend the ideas of DLA for handwritten text documents to perform region segmentation, region classification and baseline detection over handwritten musical scores in an integrated manner. Several experiments were carried out on two different datasets in order to validate this approach and assess it in different scenarios. Results show high accuracy in such complex manuscripts and very competent computational time, which is a good indicator of the scalability of the method for very large collections.

Keywords: document layout analysis · text region detection and classification · semantic segmentation · music document processing · music score images

1 Introduction

Thousands of handwritten documents are available in libraries and other institutions around the world, but most of them are not searchable or even browsable by modern digital means due to the lack of available digital transcripts.

Music constitutes one of the main vehicles for cultural transmission, hence handwritten musical scores have been preserved over the centuries, thus it is important that they can be studied, analyzed, and performed.

Handwritten musical scores are very complex and because of the huge amount of documents available it is intractable to provide accurate transcripts in a totally manual manner. Consequently, automatic or semi-automatic transcription systems have been developed to accelerate this process.

Those systems are often divided into two sub-process: Handwritten Text Recognition (HTR) [12, 14] for the lyrics and other textual regions of the document, and Handwritten Music Recognition (HMR) [2, 5] for the musical content of the document (e.g. the staff).

Therefore, before attempting to recognize the content written in a musical document, we must perform document layout analysis on it, i.e. divide it into relevant regions which can be processed by HTR, HMR systems or any other system available for a specific region. Those regions must be physically segmented from the document (region segmentation) and labeled accordingly (region classification).

In this work, we present a system based on Artificial Neural Networks, which is able to segment the document into relevant regions, provide the label associated to each region and detect the baselines (the imaginary lines upon which the lines of text rest) on text regions. It is an integrated approach where regions and baselines are segmented and detected in a single process.

The rest of the paper is organized as follows: first in Section 2 we present the current state of the art regarding music layout analysis. Section 3 provides an overview of the layout analysis technologies used. In Section 4 we present in detail the corpora used in the experiments, the evaluation measures and the system set-up. Then, the results are presented in Section 5. Section 6 closes the paper with the conclusions.

2 Related Work

Although the importance of Document Layout analysis (DLA) process before any recognition step is clear, only a few studies have tackled the task for handwritten music scores. Most of them have focused on pixel-wise classification of the different symbols or elements present in the staff (staff line, symbol, text) [3, 4, 7, 13] where the staff itself is supposed to be segmented previously or later by another method.

Other methods focus on separating music and lyrics sections, by searching local minima points on binarized images [1] or using projection profiles and Hidden Markov Models [6], but they are restricted to documents where all regions follow a vertical order (no horizontal split is allowed).

Approaches based on detailed pixel-level classification or symbol level classification are not scalable due to the cumbersome process of ground-truth generation. Also, approaches restricted by the vertical order of the document fails in many complex scenarios. For this reason we propose an integrated approach which is able to separate the different regions in a document using a region-level ground-truth instead of pixel-level without any vertical nor horizontal restriction. It follows the ideas previously introduced for DLA in handwritten text images [11].

3 Framework Description

This works extents on the ideas successfully applied to Handwritten Document Layout Analysis [11, 12]. Here we show the applicability of those methods to the complex task of Layout Analysis on handwritten music scores.

Following similar formulation as in [11], DLA on music documents is defined as a two task problem:

- *Task-1*: Region segmentation and labeling.
- *Task-2*: Baseline detection.

Task-1 consists in classifying the input image into a set of regions and assign each one to the correct class (e.g. heading, paragraph, staff, etc). On the other hand, *Task-2* consists in obtaining the baseline of each text line present in the regions where text is expected (e.g. heading, paragraph).

The proposed method consists of a set of two main stages used to solved the multi-task problem formulated previously in an integrated manner[1]. In the first stage (called Pixel level classification) an Artificial Neural Network (ANN) is used to classify the pixels of the input image ($\boldsymbol{x} \in \mathbb{R}^{w \times h \times \gamma}$, with height h, width w and γ channels) into a defined number of zones of interest (text, illustration, staff, etc) and baselines. In the second stage (called Zone segmentation and baseline detection) a contour extraction algorithm is used to consolidate the pixel level classification into a set of simplified regions delimited by closed polygons. Then a similar process is carried inside each region where a line of text is expected to extract the baselines. In contrast to [**?**], here we restrict the search of baselines to only those regions where text is expected to be (e.g. a region of type "lyrics" is expected to contain text while "staff" does not).

Stage 1, pixel level classification given an input image x, we can define a multi-task variable [2] $\boldsymbol{y} = [\boldsymbol{y}^1, \boldsymbol{y}^2]$, where $\boldsymbol{y}^t = (y_{ij}^t), 1 \leq i \leq w, 1 \leq j \leq h, 1 \leq t \leq 2$ and $y_{ij}^t \in \{1, \ldots, K^t\}^{w \times h}$ with $K^t \in \mathbb{N}_+$ being the finite number of classes associated with the t-th task. The solution of this problem for some test instance $\boldsymbol{x}$ is given as the following optimization problem:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}) \tag{1}$$

where the conditional distribution $\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x})$ is usually unknown and has to be estimated from training data $D = \{(\boldsymbol{x}, \boldsymbol{y})\}_{n=1}^{N} = \{(\boldsymbol{X}, \boldsymbol{Y})\}$.

In our case *Task-2* ($t = 2$) is a binary classification problem, then $K^2 = 2$ (background, baseline), and *Task-1* ($t = 1$) is a multi-class problem where K^1 is equal to the number of different types of regions in the specific corpus, plus one for the background.

In this work the conditional distribution $\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x})$ is estimated under naive Bayes assumption for each pixel in the image by *M-net*, the Conditional Adversarial Network presented in [11]. Under this assumption the optimization problem formulated in Eq. (1) can be computed in exact manner element by element, this is:

[1] Notice that both stages works together on both tasks.

[2] For convenience, each task will be represented mathematically as a superscript over the variables (e.g. v^t).

– *Task-1*:

$$y_{i,j}^{*1} = \underset{y \in \{1,\ldots,K^2\}}{\arg\max} \mathcal{M}_{i,j,y}(\boldsymbol{x}), \quad 0 \leq i \leq w, 0 \leq j \leq h \tag{2}$$

– *Task-2*:

$$y_{i,j}^{*2} = \underset{y \in \{0,1\}}{\arg\max} \mathcal{M}_{i,j,y}(\boldsymbol{x}), \quad 0 \leq i \leq w, 0 \leq j \leq h \tag{3}$$

where $\mathcal{M}(\cdot)$ is the output of the latest layer of *M-net*.

Stage 2, Zone segmentation and baseline detection algorithm let a test instance $\boldsymbol{x}$ and its pixel level classification $\boldsymbol{y}^*$ obtained in the previous stage be given. First, the contour extraction algorithm presented by Suzuki et al. [15] is used for each region type over $\boldsymbol{y}^1$ to determine the vertices of its contour (we call it a region-contour). Then, for each region-contour found, which belongs to a region where a line of text is expected, we apply the same extraction algorithm over $\boldsymbol{y}^2$ to find the contours where baselines are expected to be (baseline-contour), but restricted to the area defined by the region polygon.

Finally, the baseline detection algorithm presented in [11] is used to detect the baseline of each baseline-contour found.

4 Experimental Setup

4.1 Corpus

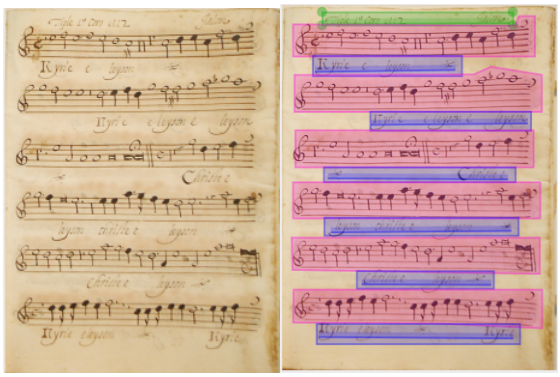
CAPITÁN is a huge archive of manuscripts of Spanish and Latin American music from the 16-th to 18-th centuries. These manuscripts were written using the so-called *white mensuralnotation*, which in many aspects differ from the modern Western musical notation. Furthermore, this archive was written following the slightly different Hispanic notation of that time, increasing its historical and musicological interest. The archive is managed by the Department of Musicology of the Spanish National Research Council of Barcelona, which kindly allowed the use of the archive for research purposes.

On this work we carried out our experiments on a subset of 96 pages of the archive, using 50 pages for training and 46 for test as defined in [6].

The dataset has been annotated manually into the following layout elements:

– **header**: title of the piece that might appear at the beginning of a piece (top of the first page).
– **staff**: represent those regions which contain a pentagram. This region does not contain text lines, hence no baselines.
– **lyrics**: words that are sung appear below their corresponding staff.

Main statistics of the dataset are presented Table 1, and an example of an annotated page in Fig. 1.



(a) (b)

Fig. 1: Example of page of the CAPITÁN dataset. (a) Original image. (b) Annotated layout, green=header, pink=staff, blue=lyrics. Better seeing in color.

Table 1: Main characteristics of the CAPITÁN dataset.

Zone	#Zones			#Lines		
Name	Train	Test	Total	Train	Test	Total
header	5	4	9	5	4	9
staff	300	276	576	—	—	—
lyrics	289	253	542	290	255	545

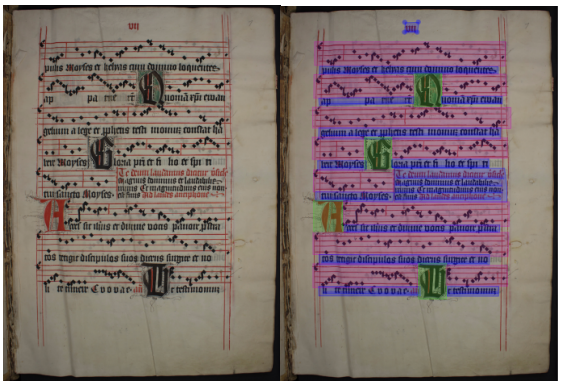
VORAU-253 is a music manuscript referred to as Cod. 253 of the Vorau Abbey library, which was provided by the Austrian Academy of Sciences. It is written in German gothic notation and dated around year 1450. This manuscript is interesting because of its complex layout. Staff, text and decorations are intertwined to compose the structure of the document (see Fig. 2 (a)).

On this work we carried out our experiments on a subset of 228 pages of the archive, using 128 randomly selected pages for training and 100 for test.

The dataset has been annotated manually into the following layout elements:

- **staff**: represent those regions which contain a pentagram. This region does not contain text lines, hence no baselines.
- **lyrics**: words that are sung appear below their corresponding staff, and other text in the document.
- **drop-capital**: a decorated letter that might appear at the beginning of a word or text line.

Main statistics of the dataset are presented Table 2, and an example of an annotated page in Fig. 2.



(a) (b)

Fig. 2: Example of page of the VORAU-253 dataset. (a) Original image. (b) Annotated layout, green=drop-capital, pink=staff, blue=lyrics. Better seeing in color.

Table 2: Main characteristics of the VORAU-253 dataset.

Zone	#Zones			#Lines		
Name	Train	Test	Total	Train	Test	Total
drop-capital	336	232	568	—	—	—
staff	1194	919	2113	—	—	—
lyrics	1379	1042	2403	1628	1215	2843

4.2 Evaluation Measures

At the best of our knowledge, there is no common evaluation measure able to assess the results obtained in both tasks jointly, therefore we present a set metrics for each task.

Task-1 Zone segmentation we report metrics from semantic segmentation and scene parsing evaluations as presented in [10]:

- Pixel accuracy (pixel acc.): $\sum_i \eta_{ii} / \sum_i \tau_i$
- Mean accuracy (mean acc.): $1/K^{t=1} \sum_i \eta_{ii}/\tau_i$
- Mean Jaccard Index (mean IU): $(1/K^{t=1}) \sum_i \eta_{ii}/(\tau_i + \sum_j \eta_{ji} - \eta_{ii})$
- Frequency weighted Jaccard Index (f.w. IU): $(\sum_\kappa \tau_\kappa)^{-1} \sum_i \tau_i \eta_{ii}/(\tau_i + \sum_j \eta_{ji} - \eta_{ii})$

where η_{ij} is the number of pixels of class i predicted to belong to class j, $K^{t=1}$ is the number of different classes for the task $t = 1$, τ_i the number of pixels of class i, and $\kappa \in \{1, \ldots, K^{t=1}\}$.

Task-2 Baseline detection we report precision (P), recall (R) and its harmonic mean (F1) measures as defined specifically for this kind of problem in [8]. Tolerance parameters are set to default values in all experiments (see [8] for details about measure definition, tolerance values and implementation details).

4.3 System Setup

Artificial Neural Network Architecture like in [11] we define *M-net* as the main network and *A-net* as the adversarial one. Both are trained in parallel, we alternate one gradient descent step on *M-net* and one step on *A-net* and so on. Both were built with 4×4 convolutional filters with stride of 2. Moreover, the architecture of each network is:

- *A-net*: `C64:C128B:C256B:C512B:C1:Sigmoid`. Activation LeakyReLU.
- *M-net*:
  - Encoder: `C64:C128B:C256B:C512B:C512B:C512B:C512B:C512`. Activation LeakyReLU.
  - Decoder: `C512BD:C512BD:C512BD:C515B:C256B:C128B:C64B:ReLU:C`$K^{t=1} + K^{t=2}$`:SoftMax`. Activation ReLU.

where `Ck` denotes a convolution layer with `k` filters, `B` a BatchNorm layer and `D` a Dropout layer with a dropout rate of 0.5, and `C`$K^{t=1} + K^{t=2}$ denotes a convolution layer with the number of filters equal to the number of output classes.

Optimization process is performed using minibatch stochastic gradient descent and Adam solver [9], with learning rate of 0.001, and momentum parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$ during 200 epochs[3]. Due memory restrictions on the hardware available, minibatch size is set to 8 images of 1024×768 on a single Titan X GPU.

Affine transformations (translation, rotation, shear, scale) and Elastic Deformations are applied to the input images as a data augmentation technique, where its parameters are selected randomly from a restricted set of allowed values, and applied on each epoch and image with a probability of 0.5. The source code used to run all experiments is available online at `https://github.com/lquirosd/P2PaLA/releases/tag/v0.6`.

5 Results

In this section we evaluate the performance of the DLA proposed approach. Experiments on each corpus were performed from very low number of training images (16) up to the maximum number of pages available for each corpus.

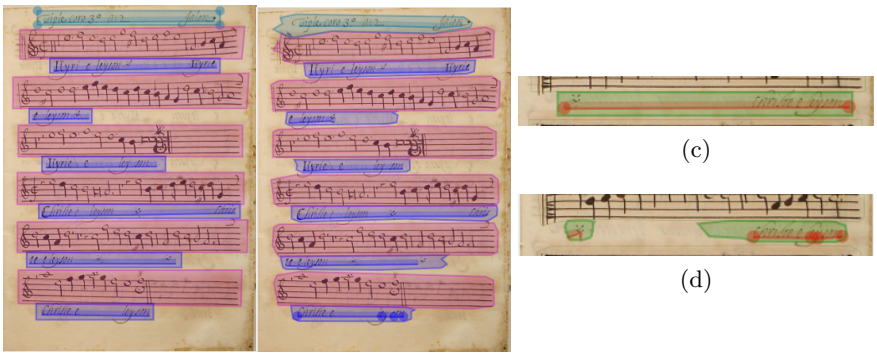
[3] In most of the cases the results can be improved by carefully select the hyperparameters of the system based on the data available for each corpus, but in this work we decide to keep hyperparameters fixed across experiments for comparability and homogeneity.

5.1 CAPITÁN

This dataset is very small, we conducted three experiments using 16, 32 and 50 pages for training, which are selected randomly and incrementally added to the training set on the respective experiment.

Task-2 (baseline) results are stable on first two experiments, but on *Task-1* an increment of around two points in most of the metrics was observed from one experiment to the next one (as expected due to the increment of training data). On the last experiment a qualitative improvement is observed, especially on baseline detection, reaching a maximum recall of 97.4% and precision of 85.0%. Errors are mainly related with zones with a long "white" space between words, as show in Fig. 3 (a,b), but because of the long space between words, the context an HTR system could use of it is negligible, hence the effect on the transcript.

In Table 3 we show the results obtained on each experiment along with the training and inference time (average in seconds per page), also we show an example of the extracted layout in Fig. 3.



(a) (b) (c) (d)

Fig. 3: Example of results obtained for the CAPITÁN test set. (a) Ground-truth image. (b) Obtained layout. (c,d) Common error observed, cyan=header, pink=staff, blue=lyrics. Better seeing in color.

We also present some results comparing our method with [6]. The definition of layout in both methods is different: on [6] is defined as a set of vertical aligned image-wide baselines for each region, and in our case we define each region by a polygon, without vertical or horizontal restriction. Consequently, we have to perform some minor adjustments on our output to make it comparable, first all regions are compressed into the horizontal line which best fits the bottom side of the polygon that defines the region, then that line is expanded up to the width of the image.

Using this simple adjustments we obtain a *relative geometric error* (RGE), as defined in [6], of 5.2% and standard deviation of 12.2%, compared to 3.2%

Table 3: Precision (P), Recall (R), F1, Pixel accuracy ($\text{Pixel}^{\text{acc.}}$), Mean Pixel accuracy ($\text{Mean}^{\text{acc.}}$), Mean Jaccard Index ($\text{Mean}^{\text{IU}}$) and Frequency weighted Jacard Index (f.w.^{IU}) results for the CAPITÁN test set. Nonparametric Bootstrapping confidence intervals at 95%, 10000 repetitions.

	Metric [%]	Number of training pages			*CI*[a]
		16	32	50	
Task-1	$\text{Pixel}^{\text{acc.}}$	90.1	91.1	92.5	±0.6
	$\text{Mean}^{\text{acc.}}$	87.7	89.6	91.4	±1.6
	Mean^{IU}	79.3	81.5	84.4	±1.8
	f.w.^{IU}	82.5	84.1	86.3	±1.0
Task-2	P	71.0	71.1	85.0	±5.4
	R	94.9	95.5	97.4	±3.5
	F1	81.2	81.5	90.8	±5.4
Training time [s]		1844.9	2641.5	3517.9	
Inference time [avg. s/page]		0.69	0.69	0.67	

[a]The confidence intervals of the elements in each row are all within the bounds listed in the corresponding row of the CI column, real values are not present in order to improve the readability of the data. Note that these intervals are not always symmetric.

and standard deviation of 3.5% in the original paper. Notice that this small difference is mostly related to the slant of the image, since we do not correct the slant, and the ground-truth used on that work is just a horizontal line.

5.2 VORAU-253

Although this dataset is far more complex than CAPITÁN, results are satisfactory even with very little training data as 16 pages, and the quality increases with the number of training data up to 87.0% mean intersection over union and 96.0% F1 measure. These results are enough to be directly processed by most of the state-of-the-art HTR and HMR systems without human intervention.

The results are show in Table 4 along with the training and inference time (average in seconds per page).An example of the extracted layout its depicted in Fig. 4.

6 Conclusions

On this paper we have proposed a new approach to undertake the DLA problem on music handwritten documents. The input image is segmented into a set of regions of interest and their baselines are detected. We have demonstrated, via empirical experiments, that the proposed approach is able to obtain a very useful and detailed layout for handwritten music documents. It can be used directly by most state-of-the art HTR and HMR systems. In addition to the encouraging

Table 4: Precision (P), Recall (R), F1, Pixel accuracy ($\text{Pixel}^{\text{acc.}}$), Mean Pixel accuracy ($\text{Mean}^{\text{acc.}}$), Mean Jaccard Index ($\text{Mean}^{\text{IU}}$) and Frequency weighted Jacard Index (f.w.^{IU}) results for the VORAU-253 test set. Nonparametric Bootstrapping confidence intervals at 95%, 10000 repetitions.

	Metric [%]	Number of training pages				*CI*[a]
		16	32	64	128	
Task-1	$\text{Pixel}^{\text{acc.}}$	92.5	93.7	94.5	94.7	±0.3
	$\text{Mean}^{\text{acc.}}$	88.6	92.0	93.4	94.2	±0.6
	Mean^{IU}	80.0	84.0	86.6	87.0	±0.7
	f.w.^{IU}	86.2	88.2	89.7	90.1	±0.5
Task-2	P	70.6	84.5	92.2	94.1	±3.0
	R	94.1	97.7	98.0	98.1	±1.2
	F1	80.7	90.6	95.0	96.0	±2.2
Training time [s]		1915.0	2814.4	4361.8	7429.8	
Inference time [avg. s/page]		1.40	1.27	1.22	1.20	

[a] The confidence intervals of the elements in each row are all within the bounds listed in the corresponding row of the CI column, real values are not present in order to improve the readability of the data. Note that these intervals are not always symmetric.

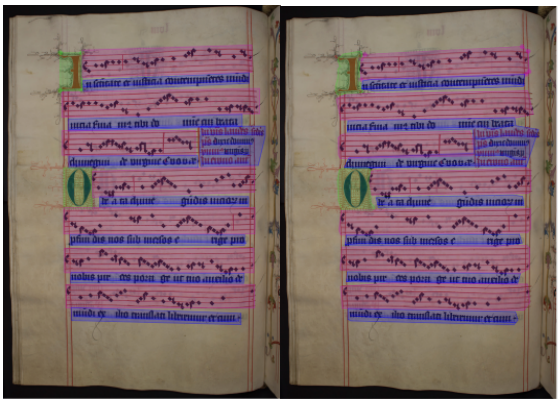
results obtained the proposed method does not rely on a very detailed labeled data, and the processing time per page is very competitive. Both characteristics are necessary for processing large scale archives. In the future we plan to extend this method to be able to extract the layout of a document in a more hierarchical way, where the relationship between regions must be extracted along with the layout.

Acknowledgment

This work was partially supported by the Universitat Politècnica de València under grant FPI-420II/899, a 2017-2018 Digital Humanities research grant of the BBVA Foundation for the project Carabela, the History Of Medieval Europe (HOME) project (Ref.: PCI2018-093122) and through the EU project READ (Horizon-2020 program, grant Ref. 674943). NVIDIA Corporation kindly donated the Titan X GPU used for this research.

References

1. Burgoyne, J.A., Ouyang, Y., Himmelman, T., Devaney, J., Pugin, L., Fujinaga, I.: 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 (2009)



(a) (b)

Fig. 4: Example of results obtained for the VORAU-253 test set. (a) Ground-truth image. (b) Obtained layout, green=drop-capital, pink=staff, blue=lyrics. Better seeing in color.

2. Calvo-Zaragoza, J., Toselli, A.H., Vidal, E.: Probabilistic music-symbol spotting in handwritten scores. In: 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 558–563 (Aug 2018). https://doi.org/10.1109/ICFHR-2018.2018.00103
3. Calvo-Zaragoza, J., Zhang, K., Saleh, Z., Vigliensoni, G., Fujinaga, I.: Music document layout analysis through machine learning and human feedback. In: 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 02, pp. 23–24 (Nov 2017). https://doi.org/10.1109/ICDAR.2017.259
4. Calvo-Zaragoza, J., Castellanos, F.J., Vigliensoni, G., Fujinaga, I.: Deep neural networks for document processing of music score images. Applied Sciences (2076-3417) **8**(5) (2018)
5. Calvo-Zaragoza, J., Toselli, A.H., Vidal, E.: Handwritten music recognition for mensural notation: Formulation, data and baseline results. In: 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 1, pp. 1081–1086. IEEE (2017)
6. Campos, V.B., Calvo-Zaragoza, J., Toselli, A.H., Ruiz, E.V.: Sheet music statistical layout analysis. In: 15th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 313–318. IEEE (2016)
7. Castellanos, F.J., J.Calvo-Zaragoza, G.Vigliensoni, Fujinaga, I.: Document analysis of music score images with selectional auto-encoders. In: 19th International Society for Music Information Retrieval Conference. pp. 256–263 (2018)
8. Grüning, T., Labahn, R., Diem, M., Kleber, F., Fiel, S.: READ-BAD: A new dataset and evaluation scheme for baseline detection in archival documents. CoRR **abs/1705.03311** (2017), http://arxiv.org/abs/1705.03311
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations (ICLR) (2015)

10. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
11. Quirós, L.: Multi-task handwritten document layout analysis. ArXiv e-prints, 1806.08852 (2018), https://arxiv.org/abs/1806.08852
12. Quirós, L., Bosch, V., Serrano, L., Toselli, A.H., Vidal, E.: From HMMs to RNNs: Computer-assisted transcription of a handwritten notarial records collection. In: 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 116–121. IEEE (Aug 2018)
13. Rebelo, A., Fujinaga, I., Paszkiewicz, F., Marcal, A.R., Guedes, C., Cardoso, J.S.: Optical music recognition: state-of-the-art and open issues. International Journal of Multimedia Information Retrieval **1**(3), 173–190 (2012)
14. Sánchez, J.A., Romero, V., Toselli, A.H., Villegas, M., Vidal, E.: ICDAR2017 competition on handwritten text recognition on the READ dataset. In: 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 1, pp. 1383–1388. IEEE (2017)
15. Suzuki, S., et al.: Topological structural analysis of digitized binary images by border following. Computer vision, graphics, and image processing **30**(1), 32–46 (1985)