Can Patch Selection Heuristics Enhance Layout Analysis of Music Scores?

Francisco J. Castellanos[0000−0001−9949−5522], Juan P. Martinez-Esteso[0009−0006−3246−3781], Alejandro Galan-Cuenca[0000−0002−5799−6270], and Antonio Javier Gallego[0000−0003−3148−6886]

University Institute for Computing Research, University of Alicante, 03690, Spain
fcastellanos@dlsi.ua.es, {juan.martinez11,a.galan}@ua.es,
jgallego@dlsi.ua.es

Abstract. Current machine-learning techniques for Layout Analysis require extensive annotated data for training. Although there are proposals to reduce the amount of annotated data, they are often limited to simple criteria for selecting regions to annotate. This work conducts a comparative study of four criteria for selecting patches to be annotated. The main objective is to analyze their influence on performance and the amount of data needed. The proposals are evaluated on four datasets, showing an average improvement of 6.7% over a random extraction method and 12% over the sequential strategy commonly used in the state of the art.

Keywords: Optical Music Recognition · Layout Analysis · Document Analysis · Deep Learning.

1 Introduction

Transcribing music documents involves extracting information from document images and converting it into a structured digital format. In contrast to storing only the scanned images of these manuscripts, this process enables searching and automated processing, thereby improving the accessibility and preservation of historical heritage. Traditionally, transcription has been performed manually, but this approach is both costly and prone to errors. In the last decade, efforts have been made to automate this task, a challenge addressed by the research field known as Optical Music Recognition (OMR).

Within the processes carried out in OMR, a crucial stage is Layout Analysis (LA), which divides music manuscripts into regions based on musical significance, such as staff lines, music symbols, or lyrics. Initial heuristic-based methods, employing techniques like the Hough transform [1] or Hidden Markov Models [7], lacked generalization and performed poorly on new documents for which they were not prepared. OMR has advanced with the application of deep learning, surpassing traditional methods in both accuracy and generalization capabilities. Deep learning-based LA approaches include pixel-by-pixel classification [2] and region segmentation [5], with the former being the focus of this work.

However, while these new techniques produce more generalized solutions, they also require a large annotated dataset for training. This is a labor-intensive task that must be performed by experts and is sometimes infeasible due to the lack of representative data, especially for old manuscripts with unique features. Data augmentation and synthetic data generation have been explored as potential solutions, but they have limitations and often produce unreliable results [8,4].

A promising solution to these challenges is the development of techniques for training models with scarce annotated datasets, known as Few-Shot Learning (FSL). For instance, the proposal by Castellanos et al. [3] remarkably reduced the amount of required annotation. However, their study was limited to a sequential criterion for selecting the subsequent areas to annotate.

In this work, we analyze alternatives to the sequential criterion, specifically comparing it with three other methods: random selection, ink rate, and entropy level. The aim is to study the influence of the selected patches in terms of performance. Experiments on four databases demonstrate the importance of the patch selection criteria. The proposal based on entropy obtains an average F_1 improvement of 12% over the sequential strategy, 6.7% over the random selection, and 1.3% over the ink rate criterion. These results suggest that more effective patch selection can lead to more robust and competitive models, reducing labeling costs by optimizing efforts. These findings could be beneficial in other FSL approaches, as well as in other areas like Active Learning or general learning-based solutions, where the data dependency is high and efficient data utilization may be beneficial.

2 Method

This section describes the sampling strategies proposed for the selection of the patches to be annotated in order to train a layout analysis learning-based model within a data scarcity scenario. For this purpose, we rely on the few-shot model established in [3]. In this proposal, competitive results were obtained with a sequential criterion for the extraction of patches. However, the appropriate choice of these patches is crucial since it can lead to better or worse results depending on the selected sample.

In this paper, we study the influence of the sample selection criterion by comparing the four alternatives described below. Note that in all cases, to avoid selecting patches that we know are useless because they contain almost no information, a subset of valid patches is created, discarding those whose percentage of ink is below a threshold (specifically, those with less than 2.5% ink).

- **Random**. The sample is chosen arbitrarily from the valid subset of patches.
- **Sequential**. The document is divided into cells with the size of the patches, from which only valid cells (those with a minimum amount of ink) are sequentially selected by rows, from left to right and top to bottom.
- **Ink-rate**. Samples are selected in descending order of the quantity of annotated pixels for a specific layer within the sample, from the highest to

the lowest within the valid subset. The ink level will depend on the layer for which the sample is selected. This strategy emulates a human criterion where samples with the most information are prioritized for selection.

- **Entropy**. Samples are selected according to the level of entropy in descending order from the set of valid patches. The entropy value is calculated using the method described in [6].

3 Experiments

In this section, we present the results obtained during the experimentation carried out to choose the best criteria for patch selection using the state-of-the-art FSL proposal by Castellanos et al. [3] for layout analysis. To conduct the quantitative evaluation, we use the F-score metric (F_1), defined as the harmonic mean of precision and recall. Experiments were performed on four datasets commonly used in the literature of OMR: Capitan, Einsiedeln, MS73, and Salzinnes [3]. These corpora include manual pixel-wise annotations of music manuscripts, categorized into four main layers of information: staff lines (*Staff*), music notation (*Notes*), lyrics (*Text*), and background (*Background*). Figure 1 includes examples of patches extracted for the *Notes* layer using the four considered criteria, with the ink ratio included. The entropy value is also displayed for this last criterion.

As a preliminary experiment, we analyze the amount of information contained in the samples extracted by each method. Figure 2 shows this analysis for the layer *Notes*, showing the ink rate according to the four criteria. As can be seen, the random and sequential methods have a similar distribution of the samples, as the ink-rate value depends on the page layout. As expected, the samples extracted with the ink-rate method follow a linear trend. Finally, the entropy method has a distinguishable pattern, where the most valuable samples are those with higher entropy values. If we compare this with the examples in Figure 1, we can notice that the entropy and ink-rate methods generally help to obtain samples with more information.

Next, we analyze the F_1 results for the layout analysis task based on the patch selection method used. Figure 3 shows this comparison according to (a) the datasets and (b) the layer of information. As we can observe in Figure 3a, the entropy-based selection proves to be the most effective for Capitan and Einsiedeln, while the ink-rate method achieves optimal results for MS73 and Salzinnes. Both methods obtain the most consistent results in the four datasets, with the entropy surpassing the average in all the datasets, and the ink rate in three of them. The sequential method only obtains positive results in Salzinnes, being outperformed in the rest of the datasets. This may be because the ink level is relatively random, only depending on the page composition. The good results for Salzinnes may be due to the fact that it contains more information on the image edges than the other datasets. Furthermore, the random method surpasses it in three datasets, which means that extracting samples randomly tends to be better in variable scenarios. These findings suggest the importance of knowing

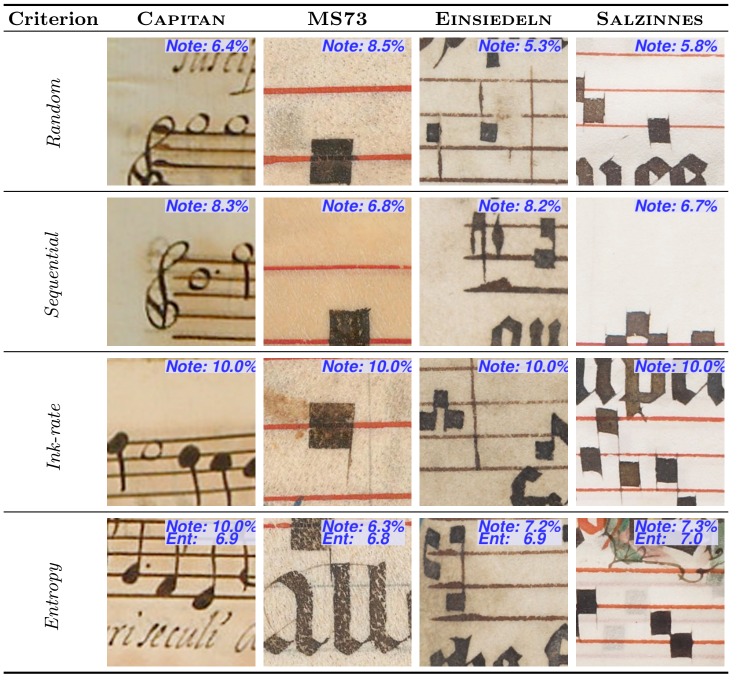
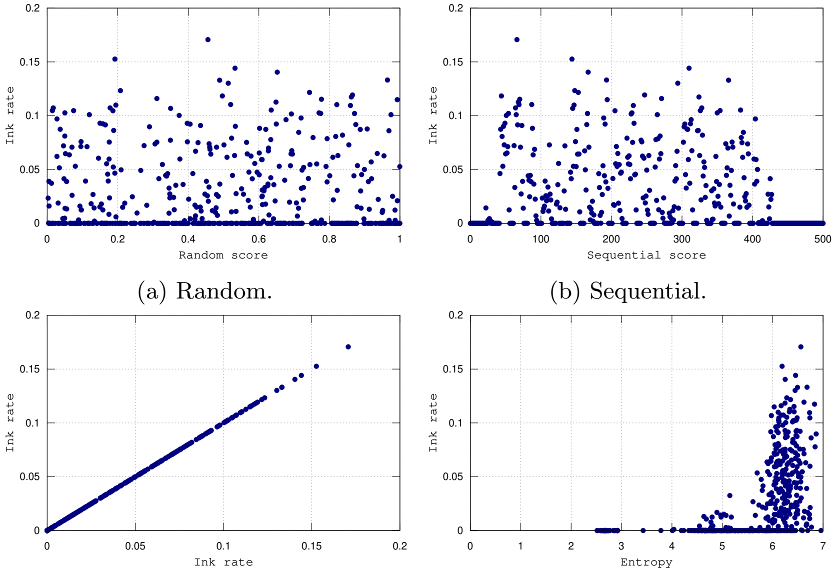


Fig. 1: Patch samples selected for annotation according to the four different criteria studied and the four corpora analyzed. These examples were extracted for the music-symbol layer. The ink percentage for this layer is shown in the upper right corner. For the entropy criterion, its value is also included.

the location of the relevant information within the image, since the entropy and ink-rate methods follow this criterion, while the random and sequential criteria may extract samples with almost no ink, leading to worse results.

For a more in-depth analysis, Figure 3b presents the results according to the layer of information. In this comparison, the entropy and ink-rate methods also achieve the best outcomes. Additionally, we can see that the *Text* and *Notes* are the hardest classes to detect, as expected due to the small size of these elements, while the lines of the *Staff* layer obtain better results as they are easier elements to detect. The high performance obtained for the *Background* class indicates that the model effectively distinguishes the relevant ink information across all datasets, keeping the number of false positives low.

Finally, the overall average results are presented in Table 1. As previously mentioned, the entropy method is the best option, closely followed by the ink-rate criteria. The random method is the least effective since it depends on the arbitrary sample selected, as does the sequential method, which depends on the composition of the document to be processed.



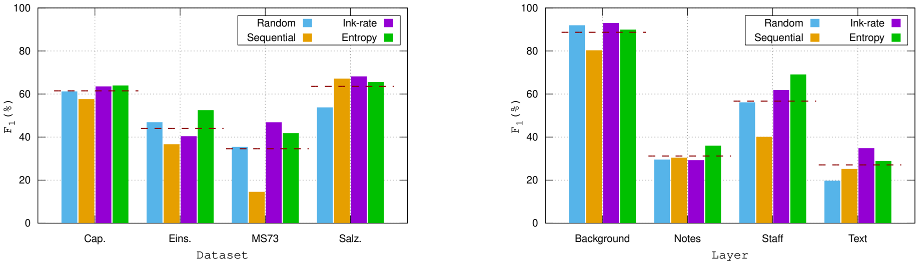
(a) Random.

(b) Sequential.

(c) Ink-rate.

(d) Entropy.

Fig. 2: Analysis of the ink rate levels in samples extracted for the *Notes* layer using the four different patch extraction methods considered.



(a) Average results per dataset.

(b) Average results per layer.

Fig. 3: Results obtained in terms of F_1 for the four patch selection methods. The dashed line in each group of columns represents the average F_1 obtained.

Table 1: Final average results in terms of F_1 for each patch extraction method. Superscript values denote the percentage improvement with respect to the random method (baseline).

	Random	**Sequential**	**Ink-rate**	**Entropy**
$\mathbf{F}_1$ (%)	49.2	$43.9^{-5.3}$	$54.6^{+5.4}$	$55.9^{+6.7}$

4 Conclusions and Future Work

This work presents a study on the influence of the criteria followed in selecting the samples to annotate in order to train a layout analysis model within a few-shot learning context. For this, four selection criteria are studied: random, sequential, ink-rate, and entropy-based. After conducting a series of experiments with four corpora, we concluded that the chosen criterion is decisive in obtaining better and more consistent results. The entropy and ink-rate-based strategies are the most effective since the samples obtained contain more valuable information for the model. On the other hand, the sequential and random methods do not consider the amount of information in each sample, resulting in lower performance by providing samples with arbitrary ink levels, depending on the page composition. Remarkably, changing to an entropy-based criterion enhances state-of-the-art results, achieving a 6.7% improvement over random selection, a 12% improvement over the sequential criterion, and even a 1.3% better over the method based on ink level. Future research aims to propose other more advanced selection methods that analyze the particular needs of the model.

Acknowledgments. This work was supported by the Spanish Ministerio de Ciencia e Innovación through the I+D+i DOREMI project (TED2021-132103A-I00), funded by MCIN/AEI/10.13039/501100011033.

Disclosure of Interests. The authors declare that they have no conflict of interest.

References

1. Burgoyne, J., Devaney, J., Ouyang, Y., Pugin, L., Himmelman, T., Fujinaga, I.: Lyric extraction and recognition on digital images of early music sources. In: 10th International Society for Music Information Retrieval Conference, ISMIR (2009)
2. Castellanos, F., Gallego, A.J., Calvo-Zaragoza, J.: Unsupervised domain adaptation for document analysis of music score images. 14th International Workshop on Machine Learning and Music (MML) (2021)
3. Castellanos, F.J., Gallego, A.J., Fujinaga, I.: A few-shot neural approach for layout analysis of music score images. In: ISMIR Hybrid Conference (2023)
4. Dey, A., Nasipuri, P.M., Das, N.: Variational augmentation for enhancing historical document image binarization. In: 13th Indian Conference on Computer Vision, Graphics and Image Processing. ICVGIP (2023)
5. Edirisooriya, S., Dong, H., McAuley, J.J., Berg-Kirkpatrick, T.: An empirical evaluation of end-to-end polyphonic optical music recognition. In: 22nd International Society for Music Information Retrieval Conference ISMIR. pp. 167–173 (2021)
6. Gray, R.M.: Entropy and Information Theory. Springer (2011)
7. Pugin, L.: Optical music recognitoin of early typographic prints using hidden markov models. In: ISMIR. pp. 53–56 (2006)
8. Tensmeyer, C., Brodie, M., Saunders, D., Martinez, T.: Generating realistic binarization data with generative adversarial networks. In: International Conference on Document Analysis and Recognition (ICDAR). pp. 172–177 (2019)