The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-M-2-2023
29th CIPA Symposium "Documenting, Understanding, Preserving Cultural Heritage:
Humanities and Digital Technologies for Shaping the Future", 25–30 June 2023, Florence, Italy

AI-ASSISTED DIGITALISATION OF HISTORICAL DOCUMENTS

S. Ferro [1,2], M. Pelillo [1,2], A. Traviglia [2,1]*

[1] Ca' Foscari University of Venice, DAIS, via Torino 155, 30172 Venice, Italy – (sara.ferro, pelillo)@unive.it
[2] Istituto Italiano di Tecnologia, Centre for Cultural Heritage Technology, via Torino 155, 30172 Venice, Italy – (Sara.Ferro, Arianna.Traviglia)@iit.it

KEY WORDS: Historical Documents, Handwriting, Digitisation, Digitalisation, Cultural Heritage, Preservation.

ABSTRACT:

Preserving historical archival heritage involves not only physical measures to safeguard these valuable texts but also providing for their digital preservation. However, merely *digitising* manuscripts and codexes is not enough. A further step is needed: the *digitalisation* of their content, i.e. the verbatim transcription of scanned texts. This process enables the accurate preservation of their textual content, making it easier to search for information and conduct further analyses. With the help of artificial intelligence, particularly Deep Neural Networks (DNNs), automatic handwriting recognition can be performed. In this study, we employed a Convolutional Recurrent Neural Network (CRNN), an established type of DNN, to determine the minimum amount of labelled data required to automatically transcribe five different historical datasets that vary in language and time period. The results show that a Character Error Rate (CER) lower than 10% can be achieved with just a few hundred labelled text lines in almost all cases.

1. INTRODUCTION

Preserving and documenting archival heritage can be achieved not only through the *digitisation* of codexes and documents (i.e., the analogue-to-digital conversion of physical items and their encoding in a digital format) but also through the *digitalisation* of their contents (i.e., verbatim transcription of scanned texts). Artificial intelligence can play a primary role in automating several processes, particularly in the area of transcribing archival documents and codexes. Machine learning models, specifically DNNs, are increasingly used for this purpose. Their performances, however, can vary significantly depending on the quality and amount of available labelled data necessary for their training. High-quality datasets of labelled data from archival documents are scarce due to the high level of domain knowledge required in the labelling process, which makes expedients like indiscriminate crowdsourced data collection not a viable option. The ability of DNNs to accurately transcribe handwritten text deteriorates with poor or limited datasets. To overcome this problem, several techniques have been developed that can be applied to the digitalisation of historical documents.
This work aims to empirically determine how many handwritten lines of historical documents written in Western languages must be manually labelled to properly train a classical DNN for Handwriting Text Recognition (HTR).

DNNs have significantly improved HTR models' performance (Lombardi and Marinai, 2020). Networks compatible with the sequential nature of handwriting – like Recurrent Neural Networks (RNNs) – and their more robust realisations (best fit for long sequences) – e.g., Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) – have been proposed. Networks that can exploit relevant information in the data, thanks to "attention mechanisms" (Bluche et al., 2017), have also been developed. However, every type of network needs enough labelled data to perform well in digitising texts. To counter this limitation, several techniques have been proposed, including data augmentation, Transfer Learning (TL) and fine-tuning (Granet et al., 2018).

Data augmentation involves increasing the number of labelled data by adding noise (e.g., salt and pepper and Gaussian noise) or performing affine transformations (i.e., translations and rotations) on the original training data to create new data. TL and fine-tuning, on the other hand, involve applying knowledge gained from solving one task (e.g., transcribing modern English handwriting) to help solve related tasks (e.g., transcribing historical documents' handwriting).

Our work aimed to use classical data augmentation techniques and fine-tuning using handwritten modern English. This type of dataset can be easily created even by non-experts to obtain a model capable of transcribing historical documents written in Latin characters. In order to evaluate the accuracy of the obtained transcription, we consider the Character Error Rate (CER) metric. A common threshold for accurately labelled text is 10% CER (Hodel et al., 2021).

2. CLASSICAL MACHINE LEARNING TRAINING TECHNIQUES

2.1 Data augmentation

Data augmentation is a technique used to increase a dataset artificially by creating modified copies of existing data. This is a common approach to enlarge the training dataset and enable the model to learn from different data than the original training set. The transformed data should be 'coherent' with the original data and represent data that could appear in reality. To create new data, basic computer vision techniques can be used, or more complex computer vision methods and deep learning models can be employed.

The classical approach to data augmentation involves simple modifications to the original data, such as affine transformations like rotations and translations or filtering the image with techniques such as blurring or colour inversion. More complex techniques involve defining potential variations of handwritten text line images. For example, Shonenkov et al. (2021) generated new data by artificially modifying handwriting with

* Corresponding author

strikethrough and creating images by combining segments cut from various samples. Generative DNNs have also been used to produce novel data with diverse styles (Gan et al., 2022).
In this work, classical data augmentation techniques such as random rotations in the range of $[-3,3]$ degrees, random vertical translations in the range of $[-0.05 * im_h, 0.05 * im_h]$, where im_h is the image height, random scaling in the range $[0.95 * orig_{scale}, orig_{scale}]$, where $orig_{scale}$ is the original scale of the image, and Gaussian blur filtering with a kernel width of 3×3 and a randomly chosen standard deviation $\sigma \in [1.0, 2.0]$ are used.

2.2 TL and fine-tuning

TL and fine-tuning are commonly employed techniques in deep learning to improve the performance of a model on a new task (Goodfellow et al., 2016). In the case of supervised learning, where data are labelled and used to teach the model to learn a task, these techniques can be especially useful when the amount of training data is limited.

The dataset of interest is typically referred to as the 'target' dataset. When implementing either TL or fine-tuning, the approach is to train the model on a distinct but 'similar' dataset, and then utilise this model as a starting point to learn a new one that is able to appropriately label the 'target' dataset.
The dataset that is used to initially train the model is called the 'source' dataset, as it is viewed as the primary source of knowledge that can be transferred to the target task. This process is known as 'pre-training', which occurs prior to any additional training that is performed on the 'target' dataset. Despite being an initial form of training, pre-training serves as a foundation for further training on the 'target' dataset. As mentioned previously, the 'source' dataset is distinct from the one that we seek to label. The 'source' dataset, nevertheless, should possess characteristics that are as similar as possible to the 'target' dataset. This is because the closer the probability distribution of the 'source' dataset is to that of the 'target' dataset, the better results are expected for labelling the 'target' dataset.
The main distinction between TL and fine-tuning lies in how the pre-trained model parameters are treated during training on new data. In TL, the model parameters that were learned through pre-training are kept fixed while training on the new data. Conversely, during fine-tuning, the entire network is trained on the new data. In both cases, new trainable layers can be added or can replace layers already present in the network to be then trained specifically for the new task.
Datasets that are commonly used for pre-training typically contain a large amount of data. In the context of image classification, notable datasets used as 'source' datasets for TL and fine-tuning include ImageNet, which comprises almost 1.3 million images and thousands of classes (Deng et al., 2009), and ImageNet-21k, which contains nearly 14 million images and 21 thousand classes (Ridnik et al., 2021). When it comes to instance segmentation, COCO (Common Objects in Context) is the preferred dataset. This dataset comprises 328 thousand images with 91 classes, as documented by Lin et al. in 2014.
Training data for historical handwriting recognition is often scarce due to various factors. One major challenge is that such datasets cannot be created using crowdsourced data collection methods. Instead, experts in Palaeography are needed to accurately label the data. Additionally, the process of digitising historical texts can be quite difficult, as they often contain noise from deterioration caused by aging, humidity, or other factors like bleed-through, creases, and scratches.

In contrast, modern English handwritten datasets are easier to transcribe and can be crowdsourced. As a result, a modern English dataset was used in this work to pre-train the model, which was then fine-tuned on historical datasets described in Section 4.

3. METRICS

The selected evaluation metric, the CER, is one of the most commonly used to measure errors in automatic handwriting recognition models.

The formula of the CER is:

$$CER = \frac{\sum_{i=1}^{n} dist_c(pred_i, true_i)}{\sum_{i=1}^{n} len_c(true_i)}, \quad (1)$$

where n is the number of samples/sequences, $dist_c(\cdot)$ is the Lenvenshtein distance calculated in characters (Levenshtein, V.I., 1966), instead $len_c(\cdot)$ is the length of the string in characters.
The Levenshtein distance is a measure of distance between two sequences: the one being evaluated and the reference one. It represents the fraction of the number of substituted, deleted and inserted elements in the sequence with respect to the number of elements in the reference/target sequence. This fraction is tipically reported as a percentage. In this study, the sequences are sequences of characters, the lines of text.

4. DATASETS

The datasets used consist of images of lines of text, which are transcribed at the line-level. The respective line-level transcription is also inputted into the model during the training.

4.1 The 'Source' Dataset

The dataset used for pre-training the network was the IAM dataset (Marti and Bunke, 2002), which is a modern English language dataset. It can be obtained from the Research Group on Computer Vision and Artificial Intelligence at the University of Bern[1] . The dataset contains 1'539 pages of scanned text from 657 different writers and a total of 13'353 isolated and labelled text lines.

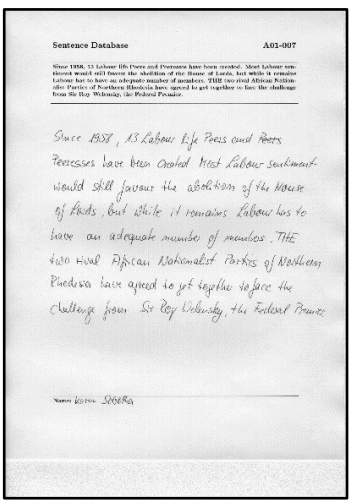

Figure 1. Page image 'a01-007x.png'.

[1] https://fki.tic.heia-fr.ch/databases

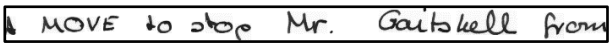

Figure 2. Sample text line image 'a01-000u-00.png'.

4.2 The 'Target' Datasets

This study utilises benchmark datasets commonly used in historical document analysis, namely the Saint Gall (Fischer et al., 2011), Parzival, and Washington datasets (Fischer et al., 2012), which, as in the case of the 'source' dataset, are available from the Computer Vision and Artificial Intelligence Research Group at the University of Bern. In addition to the benchmark datasets commonly used in historical document analysis, we also utilised an in-house created historical document, the Specchieri Marigold dataset, which comprises two subsets: the Specchieri Marigold Style 0 and the Specchieri Marigold Style 1. This dataset was created to provide a diverse set of historical documents for training and testing the model and includes documents of two different styles presenting several abbreviations. The introduction to all datasets will be presented in the following subsections.

4.2.1 The Saint Gall dataset comprises images of manuscripts written in Latin dating back to the 9th century. Figure 3 illustrates an image of a page from one of these historical documents, while Figure 4 displays a sample image of a text line.

Figure 3. Page image 'csg562-003.jpg'.

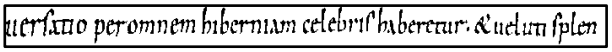

Figure 4. Sample text line image 'csg562-003-01.jpg'.

4.2.2 The Parzival dataset comprises images of 13th century manuscripts written in German. An image of a page from this historical document is depicted in Figure 5, while Figure 6 shows a sample image of a text line.

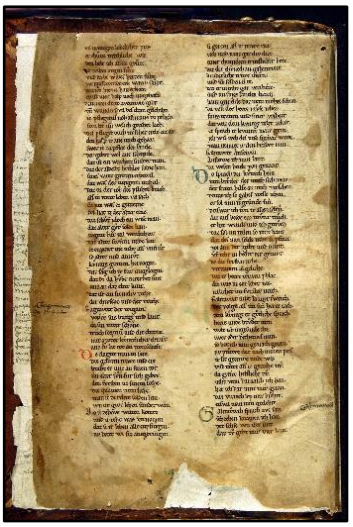

Figure 5. Page image 'd-006.jpg'.

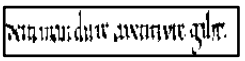

Figure 6. Sample text line image 'd-006a-001.jpg'.

4.2.3 The Washington dataset consists of pages from the George Washington Papers written in the 18th century. While an image of the document itself is not provided with the dataset, a sample image of a text line is shown in Figure 7.

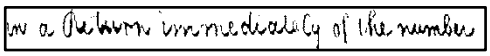

Figure 7. Sample text line image '270-27.jpg'.

4.2.4 The Specchieri Marigold dataset comprises images of 18th century writings in old Venetian language and Latin. The dataset contains images of pages from a book that reproduces older documents. It is divided into two subsets that represent different writing styles. The first subset features a precise and regular writing style, while the second subset showcases a nearly free-hand style that is similar to modern handwriting. These two subsets are named 'Specchieri Marigold Dataset Style 0' and 'Specchieri Marigold Dataset Style 1', respectively. Figure 8 displays an image of a page from 'Specchieri Marigold Dataset Style 0', and Figure 9 shows a sample image of text lines from this dataset. Figure 10 shows an image of a page from 'Specchieri Marigold Dataset Style 1', and Figure 11 displays a sample image of text lines from this dataset.

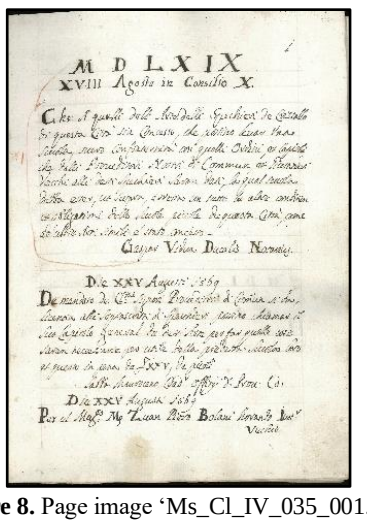

Figure 8. Page image 'Ms_Cl_IV_035_001.jpg'.

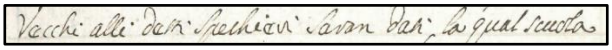

Figure 9. Sample text line image 'Ms_Cl_IV_035_001_7.png'.

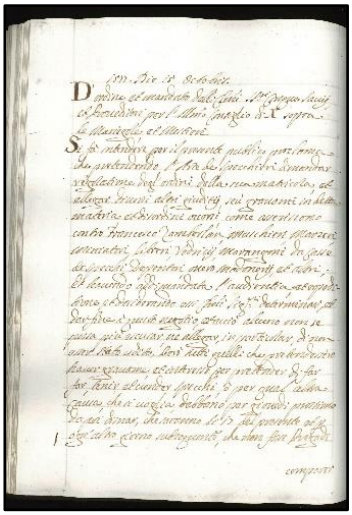

Figure 10. Page image 'Ms_Cl_IV_035_068.jpg'.

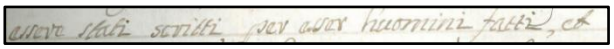

Figure 11. Sample text line image 'Ms_Cl_IV_035_025_1.png'

4.3 Partitions

The 'source' dataset is commonly used in scientific papers, but there is often uncertainty regarding the optimal partitioning of the data into training, validation, and test sets. However, a common partition includes 6'482 lines for training, 976 lines for validating, and 2'915 lines for the test set.

Table 1 presents the original size of the 'target' training dataset as provided by the authors of the datasets, along with the size of the dataset used for evaluation. The validation and test sets were combined into a single 'Evaluation' set. In order to carry out the experiments, the number of training lines was limited to a few hundred. The evaluation results were obtained at the end of the training epochs, guaranteeing that no information beyond the reduced training dataset was utilised.

Dataset Name	Training Set	'Evaluation' Set
Saint Gall dataset	468	942
Parzival dataset	2'237	2'240
Washington dataset	325	331
Specchieri Marigold dataset Style 0	635	424
Specchieri Marigold dataset Style 1	1672	719

Table 1. Cardinality of the 'target' datasets.

5. PRE-PROCESSING

The pre-processing procedure suggested by (Retsinas et al., 2022) was used to prepare the datasets for training the model. The images were first resized to a fixed height of 128 pixels and centred in a dimension of 1024 pixels of width. In cases where the original image was smaller, the borders were padded with a fixed value that is the median value of the original image. This ensured that all images used in the model have the same dimensions of 128×1024 (H×W), necessary for being inputted into the model.

6. HANDWRITING RECOGNITION MODEL

This work used a CRNN with image line-level data, a classical model for handwriting recognition. The model we used is released in (Retsinas et al., 2022) and consists of three main blocks: the first for feature extraction, the second for flattening the output of the previous module for input to the final block for sequence labelling (see Figure 10). The feature extraction block includes convolutional layers, which are widely used for feature extraction in general images since they can learn the features that produce the best recognition results (Krizhevsky, A., 2012), as well as Max Pooling layers for dimensionality reduction. The flattening block is implemented using a Max Pooling layer. Finally, sequence labelling is performed using a type of RNN, the Bidirectional LSTM (BLSTM) able to exploit left-to-right context and right-to-left one. The loss used to train the CRNN is the same as in (Retsinas et al., 2022).

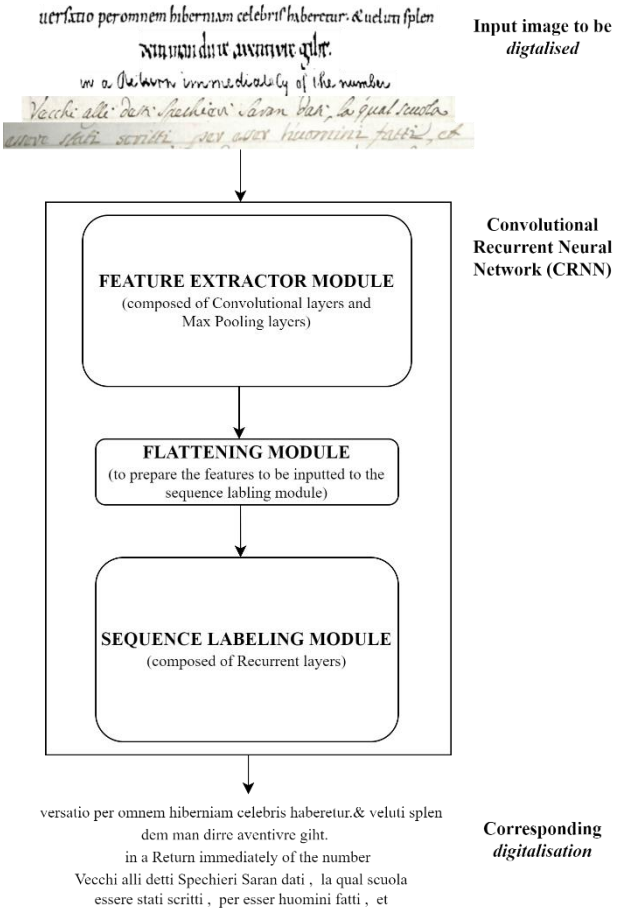

Figure 10. A high-level overview of the model.

7. SETTINGS

The model used for fine-tuning (the one trained on the 'source' modern English dataset) was obtained with the same settings as in (Retsinas et al., 2022). They used a classical optimiser, Adam, to update the CRNN parameters during training (Kingma and Ba, 2014). An initial learning rate of 1E-3, a weight decay of 5E-5, and a fixed number of epochs equal to 240 was used, while decreasing the learning rate of 0.1 at specific epoch numbers (120 and 180).

In the experiments on the historical datasets, a lower learning rate of 1E-4 was utilised to prevent overfitting, which occurs when a model becomes too dependent on the training data and is unable to generalize to new data. To fine-tune the network, we used the pre-trained model on the modern English dataset and adjusted the final layers to ensure that the probability distribution over the classes is consistent with characters/graphemes present in the alphabets of the historical datasets.

8. EXPERIMENTS

The 'target' datasets were subject to various experiments, and the results are presented in Tables 2-11. For each dataset, two tables are presented: the first one shows the CER of the 'baseline' model that does not employ either data augmentation or fine-tuning, while the second table reports the results when using either data augmentation without pre-training, or both. As expected, the CER increases as the amount of training data decreases. The table with the lowest CER among the two tables per dataset is highlighted in bold. In addition, all CERs less than or equal to 10% are highlighted in italics.

N. Training Samples	CER
100	19,53%
200	10,82%
300	*8,13%*

Table 2. Results on the Saint Gall dataset for the 'baseline' case.

N. Training Samples	CER and w/ data augmentation w/o pre-training	CER and w/ data augmentation w/ pre-training
100	*8,12%*	*9,54%*
200	*5,68%*	*6,63%*
300	***4,98%***	*5,58%*

Table 3. Results on the Saint Gall dataset w/ data augmentation and w/o pre-training, and w/ both.

N. Training Samples	CER
100	96,14%
200	23,48%
300	15,4%

Table 4. Results on the Parzival dataset for the 'baseline' case.

N. Training Samples	CER and w/ data augmentation w/o pre-training	CER and w/ data augmentation w/ pre-training
100	96,38%	65,48%
200	*6,46%*	*6,94%*
300	***2,97%***	*4,43%*

Table 5. Results on the Parzival dataset w/ data augmentation and w/o pre-training, and w/ both.

N. Training Samples	CER
100	69,39%
200	28,44%
300	23,33%

Table 6. Results on the Washington dataset for the 'baseline' case.

N. Training Samples	CER and w/ data augmentation w/o pre-training	CER and w/ data augmentation w/ pre-training
100	24,57%	46,35%
200	*9,20%*	*9,65%*
300	*6,48%*	***6,15%***

Table 7. Results on the Washington dataset w/ data augmentation and w/o pre-training, and w/ both.

N. Training Samples	CER
100	67,73%
200	34,64%
300	23,62%

Table 8. Results on the Specchieri Marigold Style 0 dataset for the 'baseline' case.

N. Training Samples	CER and w/ data augmentation w/o pre-training	CER and w/ data augmentation w/ pre-training
100	44,37%	30,8%
200	15,34%	12,27%
300	11,12%	***9,54%***

Table 9. Results on the Specchieri Marigold Style 0 dataset w/ data augmentation and w/o pre-training, and w/ both.

N. Training Samples	CER
100	65,97%
200	32,94%
300	24,77%

Table 10. Results on the Specchieri Marigold Style 1 dataset for 'baseline' case.

N. Training Samples	CER and w/ data augmentation w/o pre-training	CER and w/ data augmentation w/ pre-training
100	31,90%	27,09%
200	14,31%	12,46%
300	11,42%	**10,56%**

Table 11. Results on the Specchieri Marigold Style 1 dataset w/ data augmentation and w/o pre-training, and w/ both.

9. RESULTS

Empirical experiments have shown that using data augmentation and fine-tuning on DNNs for historical handwriting recognition can achieve a CER of less than 10% with only a few hundred lines of text from the 'target' dataset in mostly all the cases. Data augmentation improves recognition capabilities and reduces recognition error. However, fine-tuning may decrease performance for certain datasets, such as the Saint Gall and Parzival datasets, possibly because the 'source' dataset differs too much from the 'target' dataset and the amount of 'target' data is insufficient to adjust the model properly. On the other hand, leveraging both data augmentation and fine-tuning yields the best results for the Washington dataset, as well as the Specchieri Marigold Style 0 and Specchieri Marigold Style 1 datasets.

10. CONCLUSIONS

Producing accurate verbatim transcription of historical documents through digitalisation is notoriously challenging due to the high level of expertise required and obtaining sufficient labelled data is often difficult. However, this study has demonstrated that data augmentation techniques applied to the

'target' dataset and fine-tuning using modern handwriting as the 'source' dataset can be effective for historical handwriting recognition when the amount of training data is limited. This method can thus help expedite the work of experts in transcribing historical documents.

It is important to note that a recognition error of 10% may not be sufficient to achieve a proper digitalisation of the documents. Therefore, this method should be considered as an aid to experts rather than a complete solution.

In summary, the combination of classical data augmentation techniques and fine-tuning using modern handwriting as a 'source' dataset can be a useful approach for automatically digitalising historical documents. However, it is important to consider the limitations of this method, and experts should continue to play a critical role in ensuring the accuracy and quality of the transcriptions.

REFERENCES

Bluche, T., Louradour, J., Messina, R., 2017. Scan, attend and read: End-to-end handwritten paragraph recognition with mdlstm attention. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)* (Vol. 1, pp. 1050-1055). IEEE.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255).

Fischer, A., Frinken, V., Fornés, A., Bunke, H., 2011. Transcription alignment of Latin manuscripts using hidden Markov models. In *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing* (pp. 29-36).

Fischer, A., Keller, A., Frinken, V., Bunke, H., 2012. Lexicon-free handwritten word spotting using character HMMs. *Pattern recognition letters*, *33*(7), 934-942.

Gan, J., Wang, W., Leng, J., Gao, X., 2022. HiGAN+: Handwriting Imitation GAN with Disentangled Representations. *ACM Transactions on Graphics (TOG)*, *42*(1), 1-17.

Goodfellow, I., Bengio, Y., Courville, A., 2016: *Deep learning*. MIT press.

Granet, A., Morin, E., Mouchère, H., Quiniou, S., Viard-Gaudin, C., 2018. Transfer learning for handwriting recognition on historical documents. In *7th International Conference on Pattern Recognition Applications and Methods (ICPRAM)*.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Hodel, T.M., Schoch, D.S., Schneider, C., Purcell, J., 2021. General models for handwritten text recognition: feasibility and state-of-the art. German kurrent as an example. *Journal of open humanities data*, *7*(13), 1-10.

Kingma, D. P., Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krizhevsky, A., Ilya S., and Geoffrey, E. H., 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*. 25. Curran Associates, Inc.

Levenshtein, V.I., 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady* (Vol. 10, No. 8, pp. 707-710).

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13* (pp. 740-755). Springer International Publishing.

Lombardi, F., Marinai, S., 2020. Deep learning for historical document analysis and recognition—A survey. *Journal of Imaging*, *6*(10), 110.

Marti, U. V., Bunke, H., 2002. The IAM-database: an English sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, *5*, 39-46.

Retsinas, G., Sfikas, G., Gatos, B., Nikou, C., 2022. Best practices for a handwritten text recognition system. In *Document Analysis Systems: 15th IAPR International Workshop, DAS 2022, La Rochelle, France, May 22–25, 2022, Proceedings* (pp. 247-259). Cham: Springer International Publishing.

Ridnik, T., Ben-Baruch, E., Noy, A., Zelnik-Manor, L., 2021. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*.

Senior, A. W., Robinson, A. J., 1998. An off-line cursive handwriting recognition system. *IEEE transactions on pattern analysis and machine intelligence*, *20*(3), 309-321.

Shonenkov, A., Karachev, D., Novopoltsev, M., Potanin, M., Dimitrov, D., 2021. StackMix and Blot Augmentations for Handwritten Text Recognition. *arXiv preprint arXiv:2108.11667*.