Do (colored) backgrounds matter? An experiment on artificially augmented ground truth for handwritten text recognition applied to historical manuscripts

Alix Chagué, Hugo Scheithauer

Do (colored) backgrounds matter? An experiment on artificially augmented ground truth for handwritten text recognition applied to historical manuscripts

Alix Chagué[1,2,3,*,†], Hugo Scheithauer[1,3,†]

[1]ALMAnaCH, Inria, 2 rue Simone Iff, 75012 Paris, France
[2]Université de Montréal, 2900 Bd Édouard-Montpetit, Montréal, QC H3T 1J4, Canada
[3]École Pratique des Hautes Études, Les Patios Saint-Jacques, 4-14 rue Ferrus, 75014 Paris, France

Abstract

We present an experiment conducted on the augmentation of older grayscale datasets designed for automatic text recognition on contemporary handwriting (IAM-Database). The augmentation method relies on the addition of colored backgrounds taken from real-world historical blank pages and allows us to create an enhanced version of IAM-Database. We train various transcription models playing on the composition of trainset and validationset using the original and enhanced IAM-Database. We test the resulting models against the original and enhanced testsets, as well as a testset composed from real-world historical documents. We find that though the transcription engine proves robust to color changes, this technique could be used to bring up to speed older grayscale datasets to create transcription models efficient on historical handwriting. Additionally, we consider the environmental costs of using enhanced data as opposed to the original dataset, and find that the impact is minor.

Keywords

Handwritten Text Recognition, Data augmentation, Low tech strategy, Historical documents, Kraken

1. Introduction

Handwritten Text Recognition (HTR) is a supervised machine learning task presenting exceptional potential for research in the humanities thanks to its capacity to open access to large amounts of primary sources. However, HTR relies on annotated data that are costly to produce and represent a challenge for the development of the technology. While more and more datasets are created to increase the performances of transcription models on historical documents, we examine the possibility to 1) bring up to speed older datasets by artificially enhancing them and 2) test their usefulness to train transcription models capable of handling historical documents.

2. Related works

Initiatives exist to ease access to a greater diversity of annotated data ready to be used to train HTR models efficient on historical documents (across languages and scripts) [1, 2]. Other datasets are specialized on a single handwriting or source [3, 4, 5]. Additionally, recent experiments offered to artificially create synthetic data resembling historical documents [6, 7] or automatically aligned transcriptions [8].

Some datasets created before the introduction of CNNs[1] are distributed with binarized/grayscale images: for example, IAM-Database [10].[2] While this dataset is often used to benchmark HTR systems [11], it is an older dataset not designed for historical documents.

An attempt to restore colors in digital copies of microfilms of medieval manuscripts did not significantly improve text recognition [12] as the HTR engine [13] proved robust to color changes.

CSDH/SCHN Congress 2024: Sustaining Shared Futures, 16–19 June 2024, McGill University, Montreal, Canada
*Corresponding author.
†These authors contributed equally.
alix.chague@inria.fr (A. Chagué); hugo.scheithauer@inria.fr (H. Scheithauer)
0000-0002-0136-4434 (A. Chagué); 0000-0002-5659-4675 (H. Scheithauer)

[1]Convolutional Neural Network, successfully introduced in HTR tasks at the beginning of the 2010s [9, p. 13], progressively removed the need for image binarization. Binarization is a preprocessing step wherein each pixel is classified between two classes: text (usually in black), or background (usually in white).
[2]It contains a total of 13,353 annotated text lines written by 657 writers between 1999 and 2002.

3. Contribution

We propose to investigate a technique to enhance IAM-Database by adding realistic colored historical backgrounds and assess the impact when applying the resulting models to real historical manuscripts.

Our contributions are:

- a new method to artificially combine colorless text line images, taken from HTR datasets, with colored real blank pages;
- an open dataset of real blank pages extracted from historical manuscripts, reusable for other tasks such as page layout detection;
- the confirmation through another approach that HTR models are robust to color changes (see [12]);
- considerations on the environmental costs of using enhanced training data.

4. Method

Our approach consists in utilizing two versions of IAM-Database (Fig. 1):

1. the original grayscale text lines images,
2. the enhanced text line images, in RGBA mode.[3]

The enhancement of images consists in combining a clip taken from the image of a real blank page[4] with the image of a text line (Fig. 2).

A real historical testset is built by combining samples from two English datasets (1850-1950) containing various hands [15, 16] (Fig. 4).

Transcription models are trained with Kraken[5], using different combinations of original/modified train and validation sets before being tested on each testset (Table 1).

5. Results

We find that a model trained exclusively on the original IAM-Database does not perform well on colored images. If we inject enhanced images during training, the model's accuracy increases slightly on the artificially colored testset, but decreases on the original testset. As can be expected in the case of a zero-shot prediction, none of the models perform well on the historical testset. However, the model trained only on the original dataset yields the worst accuracy, suggesting that the colored samples may have positively impacted the models.

6. Considerations on environmental costs

Following [17]'s recommendations, we used Green Algorithms' Calculator [18][6] to evaluate the environmental impact of using the enhanced dataset against the original IAM-Database. While the size of the enhanced dataset is considerably increased (Table 2), we find that the computation time for the enhanced dataset is lower, with a positive impact on the environmental cost of training (Fig. 5).

7. Discussions

Our method opens the possibility to aggregate the enhanced IAM-Database with other real-world ground truth datasets, thus contributing to augmenting the quantity of data available to train generic transcription models.

Further investigations could be conducted on the usefulness of our method to enhance synthetically generated ground truth[7].

[3] In grayscale mode, the image contains only one layer of information (there is only one color channel) whereas in RGBA mode, the image contains 4 layers of information: three for the primary colors (Red, Green, Blue) and one for transparency (Alpha). Converting the images to RGB is necessary to obtain colored images, and the Alpha channel is necessary for our blending technique.

[4] We created a dataset of blank pages taken from random real manuscripts available in Gallica, the French national library's online library. The dataset is called Gallicalbum [14] and consists of 111 images.

[5] We use Kraken 4.3.0. with its default training architecture and a learning rate of $1e^{-4}$.

[6] Accessible at http://calculator.green-algorithms.org/

[7] Using techniques such as those presented in [19] or [20], which usually generate black texts on white backgrounds.

8. Acknowledgements

The authors are grateful to the CLEPS infrastructure from the Inria Paris Center for providing resources and support.

References

[1] A. Chagué, T. Clérice, "I'm here to fight for ground truth": HTR-United, a solution towards a common for HTR training data, in: Digital Humanities 2023: Collaboration as Opportunity, 2023.

[2] A. Pinche, T. Clérice, A. Chagué, J.-B. Camps, M. Vlachou-Efstathiou, M. G. Levenson, O. Brisville-Fertin, F. Boschetti, F. Fischer, M. Gervers, A. Boutreux, A. Manton, S. Gabay, P. O'Connor, W. Haverals, M. Kestemont, C. Vandyck, CATMuS-Medieval: Consistent Approaches to Transcribing ManuScripts, 2023.

[3] N. Serrano, F. Castro, A. Juan, The RODRIGO Database, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, 2010.

[4] J. A. Sánchez, Bentham Dataset R0, 2016. doi:10.5281/zenodo.44519.

[5] A. Scius-Bertrand, P. Ströbel, M. Volk, T. Hodel, A. Fischer, The Bullinger Dataset: A Writer Adaptation Challenge, in: G. A. Fink, R. Jain, K. Kise, R. Zanibbi (Eds.), Document Analysis and Recognition - ICDAR 2023, Lecture Notes in Computer Science, Springer Nature Switzerland, Cham, 2023, pp. 397–410. doi:10.1007/978-3-031-41676-7_23.

[6] L. Vögtlin, M. Drazyk, V. Pondenkandath, M. Alberti, R. Ingold, Generating Synthetic Handwritten Historical Documents with OCR Constrained GANs, in: J. Lladós, D. Lopresti, S. Uchida (Eds.), Document Analysis and Recognition – ICDAR 2021, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2021, pp. 610–625. doi:10.1007/978-3-030-86334-0_40.

[7] C. Vidal-Gorène, J.-B. Camps, They're veGAN but they almost taste the same: Generating simili-manuscripts with artificial intelligence, in: Digital Humanities 2023: Collaboration as Opportunity, 2023.

[8] S. Tarride, T. Faine, M. Boillet, H. Mouchère, C. Kermorvant, Handwritten Text Recognition from Crowdsourced Annotations, 2023. doi:10.48550/arXiv.2306.10878. arXiv:2306.10878.

[9] B. Kiessling, Avancées En Reconnaissance Optique Des Caractères Pour Les Documents Arabes Historiques / Advances in Optical Character Recognition for Historical Arabic Documents, Ph.D. thesis, Ecole Pratique des Hautes Etudes, Paris, 2021.

[10] U.-V. Marti, H. Bunke, The IAM-database: An English sentence database for offline handwriting recognition, International Journal on Document Analysis and Recognition 5 (2002) 39–46. doi:10.1007/s100320200071.

[11] W. AlKendi, F. Gechter, L. Heyberger, C. Guyeux, Advancements and Challenges in Handwritten Text Recognition: A Comprehensive Survey, Journal of Imaging 10 (2024) 18. doi:10.3390/jimaging10010018.

[12] T. Clérice, A. Pinche, Artificial colorization of digitized microfilms and its impact on other tasks, 2021.

[13] B. Kiessling, M. T. Miller, R. Maxim G, S. B. Savant, Important New Developments in Arabographic Optical Character Recognition (OCR), Al-Uṣūr al-Wusṭā 25 (2017) 1–13.

[14] A. Chagué, H. Scheithauer, Gallicalbum, 2023.

[15] K. Pham, University of Denver Collections as Data - HTR Train and Validation Set JCRS_2020_5_27, 2020. doi:10.5281/zenodo.4243023.

[16] J. Schaefer, K. Ross-Jones, A. Litvine, Joseph Hooker HTR, 2023.

[17] L. Bouza Heguerte, A. Bugeau, L. Lannelongue, How to estimate carbon footprint when training deep learning models? A guide and review, Environmental Research Communications (2023). doi:10.1088/2515-7620/acf81b.

[18] L. Lannelongue, J. Grealey, M. Inouye, Green Algorithms: Quantifying the Carbon Footprint of Computation, Advanced Science 8 (2021) 2100707. doi:10.1002/advs.202100707.

[19] A. Graves, Generating Sequences With Recurrent Neural Networks, 2014. doi:10.48550/arXiv.1308.0850. arXiv:1308.0850.

[20] T. Luhman, E. Luhman, Diffusion models for Handwriting Generation, 2020. doi:10.48550/arXiv.2011.06704. arXiv:2011.06704.
[21] A. Chagué, T. Clérice, J. Norindr, M. Humeau, B. Davoury, E. V. Kote, A. Mazoue, M. Faure, S. Doat, Manu McFrench, from zero to hero: Impact of using a generic handwriting recognition model for smaller datasets, in: Digital Humanities 2023: Collaboration as Opportunity, 2023.

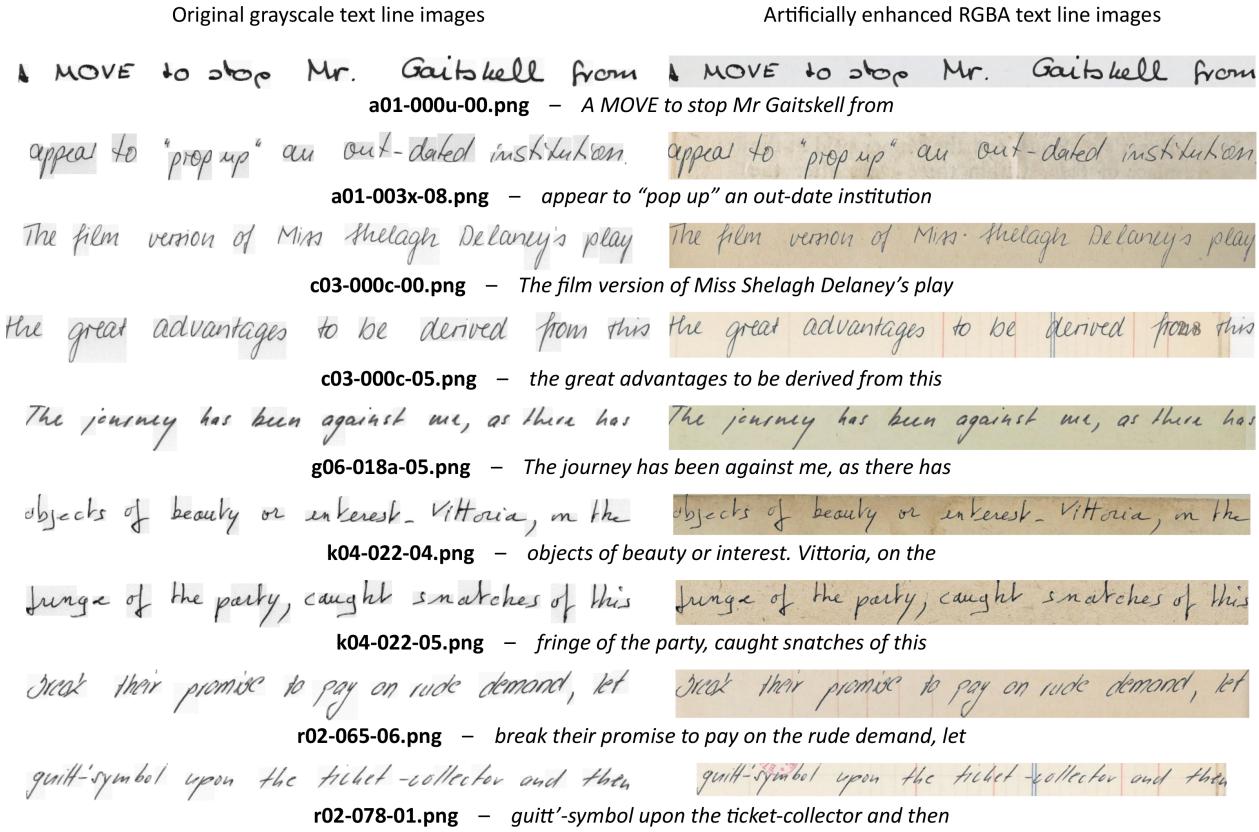

Figure 1: Samples of text line images taken from the original dataset and their equivalent in the enhanced dataset.

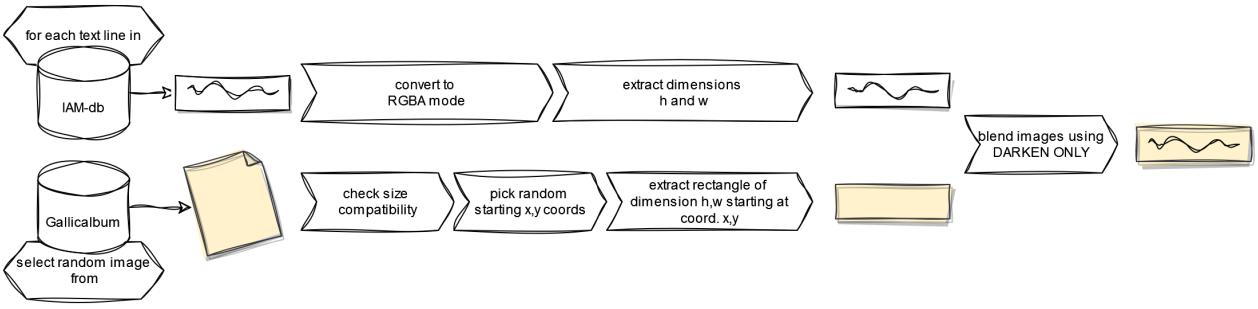

Figure 2: Modeling the workflow applied to blend real backgrounds into the text line images of the IAM-Dataset

Figure 3: Examples of blank pages collected in Gallicalbum.

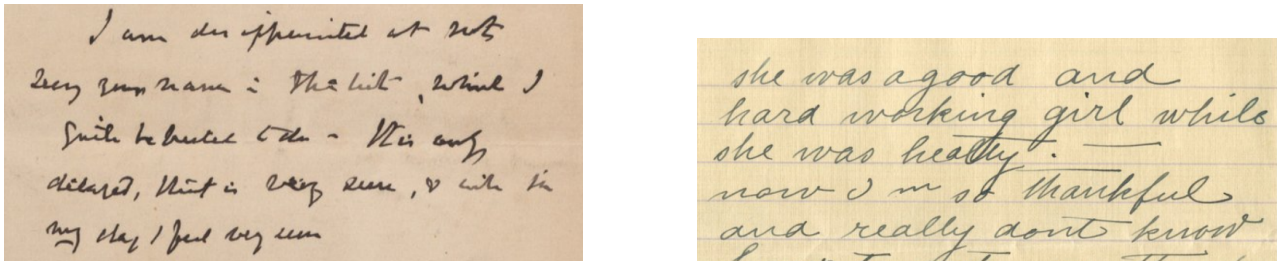

Figure 4: Two extracts of handwritten texts taken from the Joseph Hooker HTR dataset (on the left, DH_2_16_160_003.jpg) and from the University of Denver Jewish Consumptives Relief Society Medical Records Validation Set (on the right, B002_01_0104_0102_004.png), illustrating the types of historical handwriting the transcription models are tested on.

Trainset	Validationset	Original* IAM testset	Enhanced** IAM testset	Historical testset
Original	Original	80.76	37.18	4.10
Original	Enhanced	76.76	67.57	17.21
Enhanced	Original	78.88	70.72	12.11
Enhanced	Enhanced	61.75	72.99	11.74
Manu McFrench Model V1		65.99	52.44	59.58

Table 1
Character accuracy measured on each testset depending on the combination of trainset and validationset. We always test the model after 100 epochs of training. Manu McFrench Model V1[21], trained on real-world data, is used for comparison. **Original=Original Grayscale mode **Enhanced=RGBA mode with blended background.*

Split	Original IAM-Database	Enhanced IAM-Database
Trainset	325 Mb	1960 Mb
Validationset	52 Mb	287 Mb
Testset	107 Mb	611 Mb
Total	484 Mb	2858 Mb

Table 2
Artificially enhancing the original grayscale dataset increased its size by almost a factor of 6.

Step	Original IAM-Database	Enhanced IAM-Database
Compiling	87s.	134s.
Training	6113s.	5160s.
Total	6200s.	5294s.
Equiv.	*1h43m20s.*	*1h28m14s.*

Table 3
Estimated computation times, including compiling arrow files and training transcription models on the corresponding arrow files.

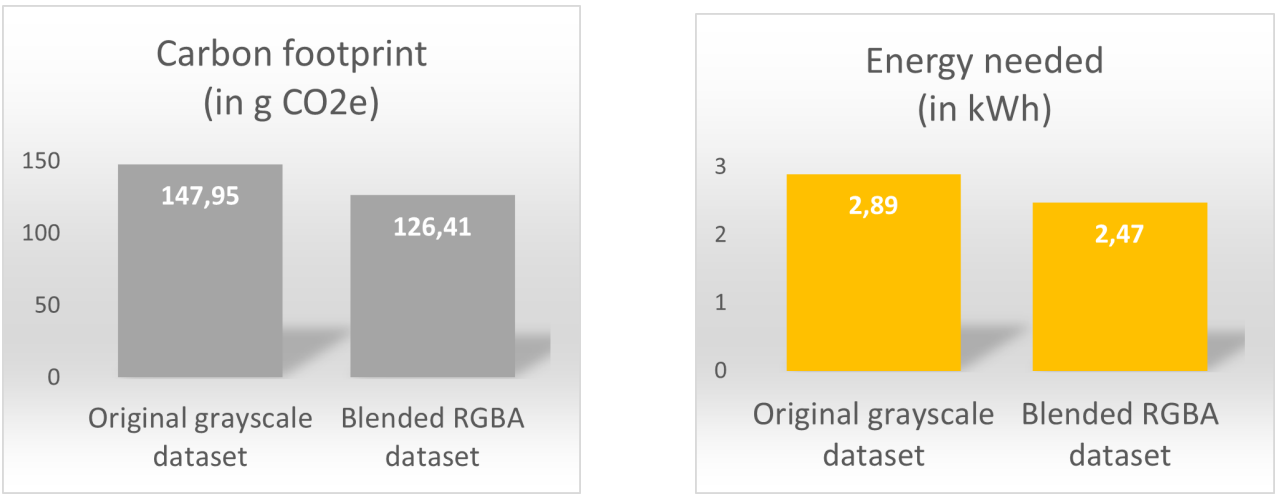

Figure 5: Comparing the impact of enhancing the dataset with our technique considering our carbon footprint and energy consumption, as estimated with Green Algorithms' Calculator on January 26, 2024.

Field	Original	Enhanced
Runtime (HH:MM)	01:43	01:28
Type of cores	Both	
Number of CPU cores	16	
Model of CPU	Other	
Thermal Design Power (TDP) value per core (CPU)	9.68	
Number of GPUs	3	
Model	Other	
TDP value per core (GPU)	260	
Memory available (in GB)	192	
Platform used for the computations	Local server	
Server location	Europe, France	
Known real usage factor of GPU	No	
Power Usage Efficiency (PUE) of data center	No	
Pragmatic Scaling Factor	No	
Carbon footprint (g CO_2e)	147.95	126.41
Energy needed (kWh)	2.89	2.47

Table 4
Details of the values passed to Green Algorithms' calculator to obtain the estimated carbon footprint and energy requirements.