



HAL
open science

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

Alix Chagué, Hugo Scheithauer

► To cite this version:

Alix Chagué, Hugo Scheithauer. Do (colored) backgrounds matter? An experiment on artificially augmented ground truth for handwritten text recognition applied to historical manuscripts. CSDH/SCHN 2024: Sustaining Shared Futures, CSDH/SCHN, Jun 2024, Montréal, Canada. hal-04622805

HAL Id: hal-04622805

<https://inria.hal.science/hal-04622805v1>

Submitted on 24 Jun 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Do (colored) backgrounds matter?

of a variable day depth. The effect of day depths upon the surfacing of various animals has been reviewed elsewhere. Anomalies. In spite of the variability of migrational behaviour, some kinds of anomalies may be recognized. Vertical movement occurs in some forms.

of a variable day depth. The effect of day depths upon the surfacing of various animals has been reviewed elsewhere. Anomalies. In spite of the variability

An experiment on artificially augmented ground truth for handwritten text recognition applied to historical manuscripts

Alix Chagué (Inria, UdeM, EPHE)
Hugo Scheithauer (Inria, EPHE)

Handwritten Text Recognition (HTR)

- supervised machine learning
- retrieve text from manuscripts
- requires annotated data
- data production is a challenge
 - variety of scripts
 - homogeneity of annotations
 - availability and findability
- reusing datasets is important

TEXT



recognition

MODEL



training



DATA

The IAM Dataset

We used the IAM Dataset for our experiment:

- created by the IAM institute at the University of Bern between 1999 and 2002
- 13K annotated text lines written in English
- 657 different writers
- distributed in the form of words, text lines, sentences or pages
- black ink on white background (grayscale images)

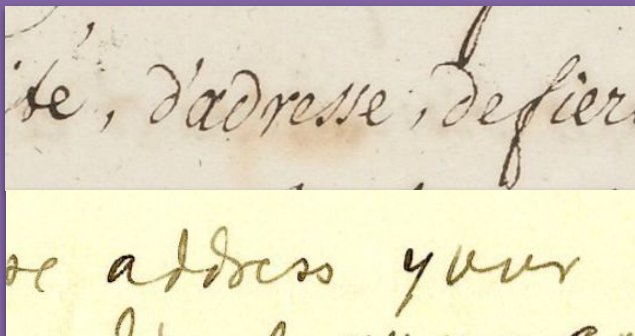
Coldly that a continuous
ge was fully place
ulation at the surface
deeper water. A
whose normal ver
on is sufficiently
to make daybird
ence at the surf
as anomalous is
siacea There are n

The IAM Dataset & historical documents



don't address me

IAM Dataset



se, d'adresse, defier

oe address your

18th / 20th centuries

- can we make datasets like IAM look more like historical documents?
- binarization is no longer necessary and recent historical document digitizations are rarely black and white
- can we upgrade the IAM Dataset instead of downgrading historical datasets? (artificial colorization/decolorization of the images)
- is it just a question of color or are the accidents in the background also meaningful?

Our questions and contribution

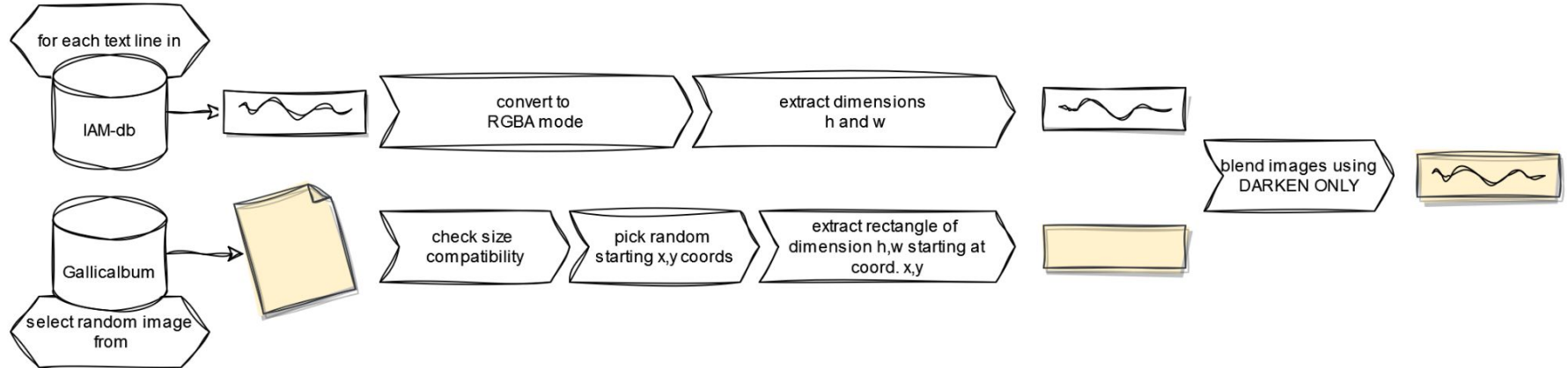
- can we make the IAM Dataset useful for historical HTR?
- are colored text lines more efficient to train HTR models rather than grayscale images?
- is it more environmentally cost-effective to use poorer grayscale images rather than rich colored images?

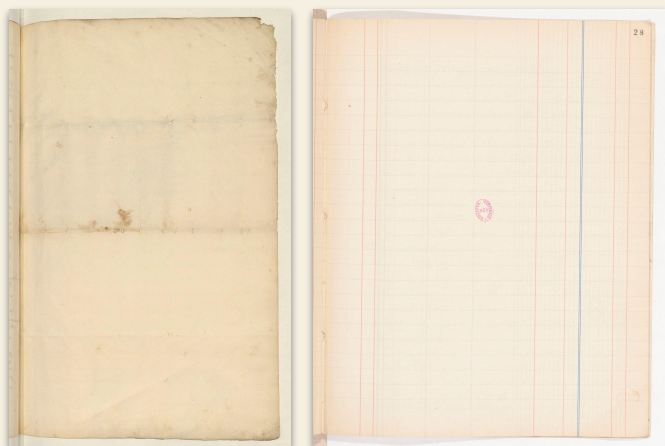
Additionally:

- we proposed a simple image manipulation method to add colored backgrounds
- we created an open dataset of real blank page documents

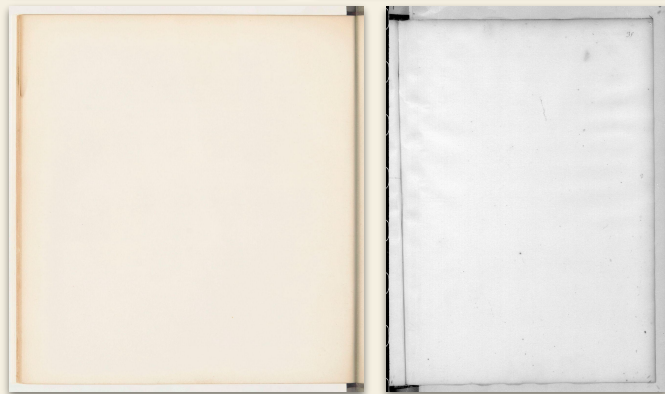
Our method

- augment a dataset with a simple image manipulation: add a richer realistic background
- train model on the original / enhanced IAM Dataset
- test resulting models on the IAM Dataset test set
- further test resulting models on real historical data





Dataset of real blank pages: Gallicalbum



- we create a dataset of real blank pages, found in Gallica
- 111 blank images, in color or not
- freely accessible: github.com/HugoSchtr/Gallicalbum
- useful to artificially colorize text images
- useful for other tasks like Layout Detection (no layout)

Enhanced IAM-Dataset

A MOVE to stop Mr. Gaitskell from A MOVE to stop Mr. Gaitskell from

a01-000u-00.png - A MOVE to stop Mr Gaitskell from

appeal to "prop up" an out-dated institution. appeal to "prop up" an out-dated institution.

a01-003x-08.png - appear to "pop up" an out-date institution

The film version of Miss Shelagh Delaney's play The film version of Miss Shelagh Delaney's play

c03-000c-00.png - The film version of Miss Shelagh Delaney's play

the great advantages to be derived from this the great advantages to be derived from this

c03-000c-05.png - the great advantages to be derived from this

The journey has been against me, as there has The journey has been against me, as there has

g06-018a-05.png - The journey has been against me, as there has

objects of beauty or interest - Vittoria, on the objects of beauty or interest - Vittoria, on the

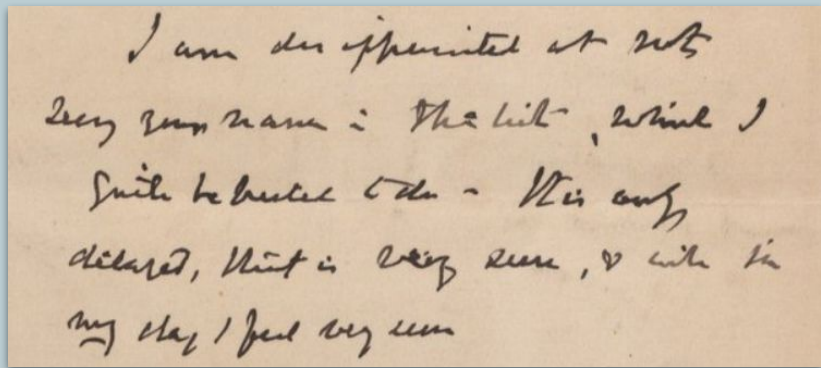
k04-022-04.png - objects of beauty or interest. Vittoria, on the

fringe of the party, caught snatches of this fringe of the party, caught snatches of this

k04-022-05.png - fringe of the party, caught snatches of this

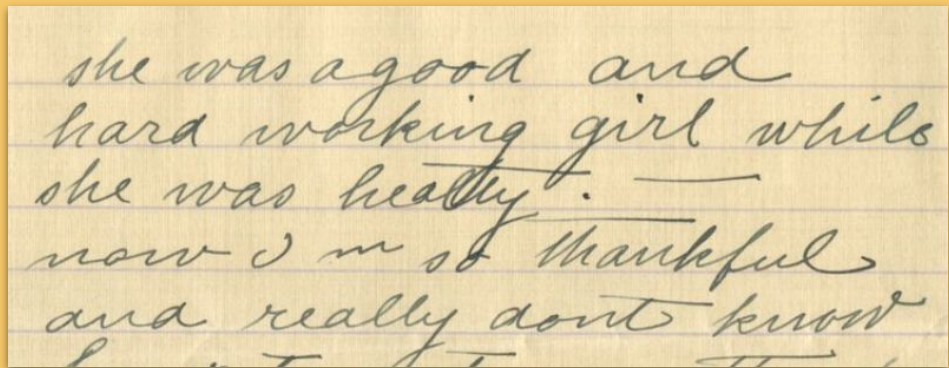
Testing on real historical data

- 2 historical datasets in English covering a period from 1850 to 1950
- listed in the HTR-United catalog and reusable (licence)
- annotation rules compatible with IAM's



I am disappointed at not
seeing your name in the list, which I
think he wanted to do - It is only
declared, that is very soon, & when in
my day I feel very soon

Joseph Hooker HTR (1850-1911)



she was a good and
hard working girl while
she was healthy. —
now I am so thankful
and really don't know

Univ. of Denver Jewish Consumptives Relief Society
Medical Records (1900-1950)

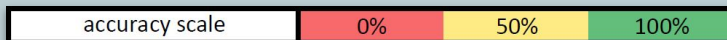
Experimentation set-up

- Kraken v 4.3.0
- default training architecture
 - 120 pixels resize, 3 CNNs + 3Bi-LSTM + 1 linear layer
- learning rate of $1e-4$ and batch size of 1
- two configs for the number of training epochs:
 - fixed (100)
 - early stopping (stops when learning curves reaches a plateau)
- train / validation / test follows the official IAM split
- test different combinations of original/enhanced train set and validation sets



Set	N text lines	N writers
train	6161	283
val 1	900	46
val 2	940	43
test	1861	128
Total	9862	500

Train set	Validation set	N epochs	IAM grayscale test set	IAM enhanced test set	Historical test set	Training time
IAM grayscale	IAM grayscale	100	69,8	29,7	4,1	01:14:33
IAM grayscale	IAM grayscale	87	70,1	30,6	7,6	01:10:19
IAM grayscale	IAM Enhanced	100	69,4	26,5	0,3	01:14:58
IAM grayscale	IAM Enhanced	62	66,4	27,9	11,1	00:51:30
IAM Enhanced	IAM grayscale	100	47,8	60,2	-4,9	01:16:52
IAM Enhanced	IAM grayscale	72	44,1	58,7	11,0	01:04:54
IAM Enhanced	IAM Enhanced	100	51,0	62,1	8,3	01:15:30
IAM Enhanced	IAM Enhanced	194	52,0	64,9	7,1	02:40:32
Manu McFrench Model V1		/	65,6	52,4	59,6	/
Manu McFondue		/	63,5	54,9	60,4	/



Our results

Discussions of the results

- probably not enough data to train an efficient model out of domain with the chosen architecture
- Manu McFrench and Manu McFondue are much better, but not good enough in zero-shot scenario on the historical testset
- models trained only on the original dataset = significant loss of accuracy on colored images (-40pts)
- models trained on colored images = minor loss of accuracy on original dataset (-10pts)
- colorful background seem to have an impact during prediction and not during training
- adding a colored background might augment the minimum amount of training data required

Perspectives to push the experiment further

- increase the quantity of data
- control the quality of the annotation in the historical test sets
- test with other historical datasets
- increase the batch size during training
- mix grayscale data and enhanced data in the train and validation sets

- Could we reach more stable scores when playing on these parameters?
- Can we reach 90% of accuracy on the IAM test set?

Final discussion on the environmental costs

- artificially enhancing IAM dataset increased its size by a factor of 6
 - Train set: 325 Mb → 1960 Mb
 - Validation set: 52 Mb → 287 Mb
 - Test set: 107 Mb → 611 Mb
- minor impact on compilation time but not on training time (fixed epoch)
- estimations from Green Algorithms' Calculator (calculator.green-algorithms.org):

	Compilation time	Training time	Total time	Carbon footprint (CO2e)	Energy needed (kWH)
All grayscale	1m27s	1h14m33s	1h16m0s	109,17	2,13
All enhanced	2m14s	1h15m30s	1h17m44s	112,04	2,18

Key takeaways

- we created a dataset of blank pages
- we successfully applied our enhancement technique to the IAM dataset
- we need to push our experimentations further to get more meaningful results
- we can already find that colorful background add noise during prediction but not during really training
- enhancing the dataset has a limited impact on the environmental cost of training a transcription model on the IAM Dataset

Dataset references:

- 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.
- K. Pham, *University of Denver Collections as Data - HTR Train and Validation Set JCRS_2020_5_27*, 2020. doi:10.5281/zenodo.4243023.
- J. Schaefer, K. Ross-Jones, A. Litvine, *Joseph Hooker HTR*, 2023.
- A. Chagué, H. Scheithauer, *Gallicalbum*, 2023.

Thank you!

And many thanks to Jennifer Carrow for her feedback on this presentation!