

Artificial colorization of digitized microfilms

a preliminary study

Thibault Clérice Ariane Pinche

Centre Jean Mabillon, École nationale des chartes - PSL

IMC 2022, July 5

Overview

1. Introduction
2. Using AI to colorized manuscripts
3. Colorization of Manuscripts and HTR Results
4. Conclusion and further explorations

Outline

1. Introduction

2. Using AI to colorized manuscripts

3. Colorization of Manuscripts and HTR Results

4. Conclusion and further explorations

Presentation

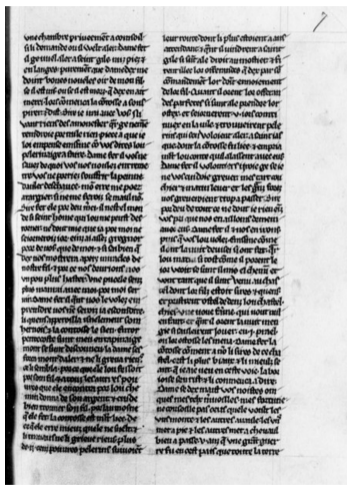
"Many low resolution digital scans of microfilms exist. These are surrogates of surrogates. They can still be (and are) profitably used, for example to corroborate a particular reading. I am however sceptical of using them as a single source for making an edition. Perhaps, indeed, 99% of a manuscript can still be deciphered by using them, but it is about that 1% of cases in which the scribe fumbled a bit with his pen and it is unclear what the word reads. In those 1% cases, you do not wish to have a low-resolution, black and white reproduction of a reproduction as your sole witness."

L. W. C. van Lit, 2019

Manuscript and online digitisation



BnF, fr.412, f.4r, 13th c.



BnF, fr. 13496, f7r, 14th c.

Short story about microfilms

- Late 1920s - early 1930s : beginning of the use of microfilm as a substitute for manuscripts
 - Facilitating the access of remote manuscripts
 - Preserving manuscripts in the context of the Second World War
- Today, cultural heritage institutions have made digitized microfilms available online through platforms such as Gallica (BnF)
 - Scanning microfilm implies lower risks
 - Lower costs than a new digitization
 - Quicker way to build a first online collection

- For some lost manuscripts, deep learning colorization is the only hope to see them again in colour.



Outline

1. Introduction

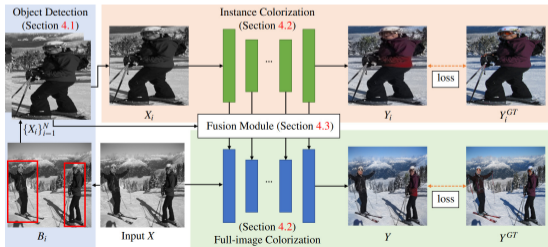
2. Using AI to colorized manuscripts

3. Colorization of Manuscripts and HTR Results

4. Conclusion and further explorations

Aims and methods

- **Aim**
 - Producing colorized images from microfilms
 - Evaluating the impact of colorization on HTR results
- **Methods**
 - Using instColorization (Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. Instance-aware image colorization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020)



Limits

- **Limits**
 - Bias can be induced by training data
 - people tend to be whitened in specific contexts
 - colourful local dresses made greyish or dulled
 - the output cannot be taken as the real colours of the original manuscript



Figure: Screenshot from <https://hyperallergic.com/639395/the-limits-of-colorization-of-historical-images-by-ai>

Colourisation Training Set

- Dataset criteria
 - Manuscripts from the 8th century to the 16th
 - Focused on west European manuscripts
 - Balanced number of representant per period
 - Representativity of the diversity of layouts
- 18 788 files from two different sources
 - Gallica or e-codices : 8660 digitized pages from 50 different manuscripts.
 - Mandragore database : 10 128 digitized pictures.



Figure: First row: random examples of pictures from the Mandragore part of the training dataset. Second row: 2 pages from composite books in the IIF part of the dataset and 3 random one from “original” manuscripts. Images display ratios are changed for the purpose of displaying multiple example on the same page.

Results



Figure: Partial reproduction of manuscripts from BnF, from the segmentation and HTR test set. Second row contains the colorized version.

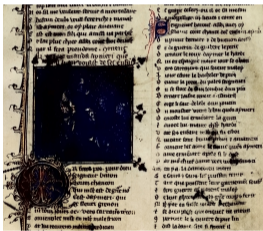
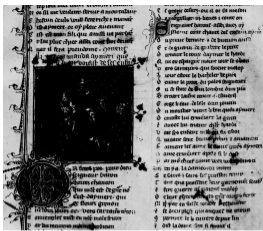


Figure: BnF fr. 24369. First row contains grayscale images used as input for *InstColorization*, second row is the corresponding output. Columns are, in order: New digitization of the manuscript, adjusted colour level of the first column to come close to microfilm contrast, digitization of the microfilm.

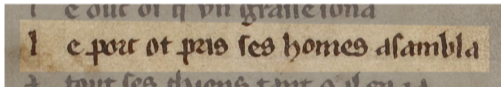
Outline

1. Introduction
2. Using AI to colorized manuscripts
- 3. Colorization of Manuscripts and HTR Results**
4. Conclusion and further explorations

Does Colourisation Affect HTR Results ?

- Our **aim** ? Finding if it helps ...
 - Segmentation recognition
 - Handwritten text recognition

Line #9



Le port ot pris ses homes asambla

by apinche (import) on Fri Apr 29 2022 10:05:58 GMT+0200



Method : tools and dataset



- Using Kraken (HTR and layout segmentation engine) for training and evaluation
- Using eScriptorium an interface for HTR ground truth production
- Using dataset based on the CREMMA Medieval dataset (Pinche, 2021)
 - Seven Old French manuscripts written between the 13th and 14th c.
 - Scanned in high definition and colour except for one manuscript
 - Sample size very different from one source manuscript to the other.

Training

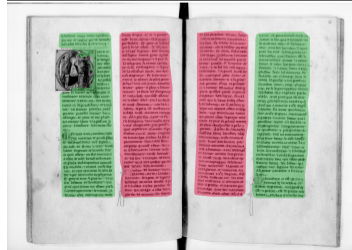
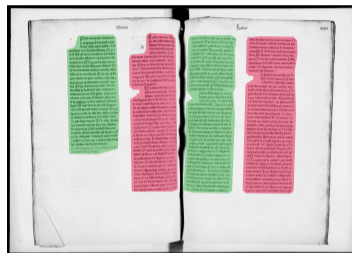
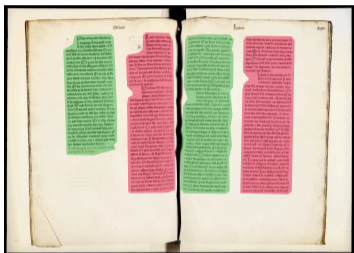
- Training two models for segmentation (one for line and one for region)
- Training two models for HTR : one with complete dataset and one without the BW document
- Using a test set from 3 different digitization of microfilm from Gallica.

Manuscript ID	Ms, fr. 13496	Ms, fr. 17229	Ms, fr. 411
Date	Late 13 th c.	Late 13 th c.– early 14 th c.	14 th c.
Annotated text	Saint Jerome's Life	Saint Lambert's Life	Saint Lambert's Life
HTR Ground Truth Locus	fol. 245v - fol.246r	fol. 163v - fol.164r	fol. 125v - fol.126r
No. of Document	1	1	1
No. of transcribed columns	4	4	4
No. of transcribed lines	159	160	150
Segmentation Ground Truth Locus	fol. 245v - fol.248r	fol. 163v - fol.169r	fol. 125v - fol.131r
No. of Document	3	6	6
No. of Segmented Columns	12	24	24

Table: Presentation of the digitized microfilmed manuscripts. Documents are double paged, composed by the verso of a folio and the recto of the following one.

Results : Zone Segmentation

Qualitatively evaluated, region segmentation is nearly not impacted by colorization.



Results : Line Segmentation

Mss, Range	Delta			Colorized Microfilm			Original Microfilm		
	Lines	M. Pixel	Pixel	Lines	M. Pixel	Pixel	Lines	M. Pixel	Pixel
fr. 411									
125v-126r	0.3	0.1	0.1	98.4	99.1	98.6	98.7	99.2	98.7
126v-127r	-0.3	0.1	0.1	100.0	99.5	99.5	99.7	99.6	99.6
127v-128r	-0.6	0.0	0.0	98.8	99.4	99.3	98.2	99.4	99.3
128v-129r	0.0	0.1	0.1	99.7	99.5	99.4	99.7	99.5	99.5
129v-130r	0.3	0.0	0.3	98.5	99.5	99.1	98.8	99.6	99.3
fr. 17229									
163v-164r	0.3	0.0	0.3	99.4	99.6	99.1	99.7	99.6	99.4
164v-165r	0.3	0.0	0.0	99.4	99.8	99.7	99.7	99.8	99.8
166v-167r	0.3	-0.1	-0.1	99.4	99.7	99.4	99.7	99.6	99.3
167v-168r	-1.0	0.0	-0.4	99.4	99.7	99.3	98.4	99.7	98.9
fr. 13496									
245v-246r	0.0	-0.0	-0.0	99.4	99.6	99.6	99.4	99.5	99.5
246v-247r	0.0	0.0	0.0	98.0	99.6	99.1	98.0	99.6	99.1
247v-248r	0.3	0.0	0.0	98.4	99.5	96.0	98.7	99.5	96.0

Results : HTR

Quantitatively evaluated, HTR results are nearly unchanged by colorization.

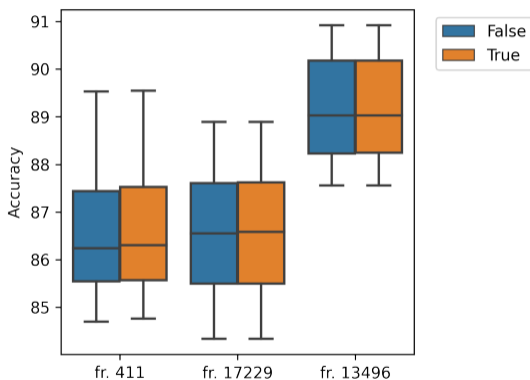


Figure: Accuracy per manuscript given the status (colorized or original gray version of microfilms).

Outline

1. Introduction
2. Using AI to colorized manuscripts
3. Colorization of Manuscripts and HTR Results
- 4. Conclusion and further explorations**

Conclusion and further explorations

- We can colorized manuscript with AI
- Artificial colorization of manuscript do not improve segmentation or HTR
- Regarding readability improvement of digitization in colour for human readers, it would be interesting to quantify the impact of artificial colorization.

Thank you for your attention !