Artificial colorization of digitized microfilms

a preliminary study

Thibault Clérice Ariane Pinche Centre jean Mabillon, École nationale des chartes - PSL

IMC 2022, July 5



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



1. Introduction

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

"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

Core Lancellon 135 g

Bassing from environment is a constant Co barfing mente (per entairerment QY d.A.sta err af is enfaigne ageiber le fabbar i la ernenafies. Cai bes gent il forma ter abing prevent a teaching and a les beffe al.

merres alm on alalent ment ber utegete hour ment piptit annere en th ferr alla Aluci 1 anne di Centraturner fi Hoursen Start and the second starter a if ether also using pirmoal truster changement and affil forfinger be defenden Len Lendennan marrie & raant mume ald me lagrant ma make charle Ce or men of met ALLOW OTHE CORE & MARTINES & DOTTOR AT A TAK OF contraste (1) and of as foreign her stated before for succlus yes commo neses enla merand hear he area and family and hearing Opier vour? uef apartie darue th mit be to mue a p must danse befen a bernef belle mene a honor . Ce apper sur m sugar tadas instant on Camblence

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

une chamber primeenter a counted Gi bemande ou d Soelr aler bame fr d genuel Alera fein ale mu piers en langes paremerare dameder me hum house nousles our bemon fil. la detturf ou le dett men. 7 ber en an ment los comence la confe A Conf ouversout day is uni ane holde Sourtene belamonether to av note venitive permile iten piece a que ie los empende emfine o voc dires los oferinaiava three Dame for & volne Cuer to quot vot not nonder entreme ne vot ne peries touthine la penne touler balebauce mo erre me poez anarment fineme forme fe maule Any for de cor bu mer d noted men be for the port of the state of the te li lente pome qui conne periri cer nomen ne tons mie que ca poe mos ne fesenerori sos envaisantes grugoror pour te not que de mos s fi entores a tor not motiven apory mundeo benofirefit-run or not beariont noo verou plat lather one puckle for the manufat anec mos por moi Cer unome ferd dur uso levoler en provider usering datum is a chemitipue internet Sperment in the chemitipue bernotige in a contrastifi le diene derore promette diene meine surresprinzige ment fehere balkoutenet is hume fait date inthespanen genetic leue fait im mit internet aussigner internet internet withen fill anzugen technistry to per-unischante die singesten zweite beer treumen die juid gentlasse ofte beer treumen die juid gentlasse ofte die fait in aussigne fait surre versite pivalore notrie Gron la chonfore. cedele erne mieur quele ne listar

teur rourse tour la plus efforent a auf aneribane som duendreur a Cum aten Girale Diviraumother th ver iller los offernites à ber sur de Smantomer lar bir emalement Telacht-Quant dosent los offerm befourferer (funrale menber for offer er feiernerver vierfanner mier en la ule provincirent och wint mui the boloiene aler: A thing is que dour la corolle futice s empris uff loucome aud statten auer euf bane fer a solomier timie or fele ne volandoie areuer met cur au Anorman town or lot Say fine uof anyuerpiene crupa pafter din pue deu de tour or ne dour le rienet SoCott and not en adlone bemen and all another d suct a upor wert a sollow usles Andine Sure don launir bruile Gon fr. A Lou maria fi coft come a posent le tor water & Generalania a Avenue wor can and fime were wider alour la file eftorr Gros vouent erputwir oftel deden loudante drief-one tione time qui nour nul entime er fin d osene tanuer men gie A Curleuren Jouer: cu-j. prnelatoff coment a no li finn be ce de Adeat is plut binne at meda a ionfe Rurifty li commouch 1 Dete Dume & der mutt vor northes om ne confortie par orut quele worth lot Surfmonnen let annet aundet die merapic Flet aunt mera cheund bien a parte Van a vac aufir ouer ry fu cuceft sait aucround la were

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.





1. Introduction

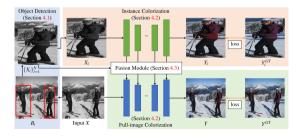
2. Using AI to colorized manuscripts

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



 $Figure: \ Screenshot \ from \ :https://hyperallergic.com/639395/the-limits-of-colorization-of-historical-images-by-aim \ for the state of the stat$

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 IIIF 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

referenden allar fellementen freid mer der wer pår genera fre geheter a	tatanter mass meren pallen. Inner depate er at fe geiner	leadar
fernen befalt men fernen teller per eller benanerei con verb i pitte		an whether the objective and the balance of the
	total and the a chief and a chief	
Council allowed on the arear Integration of the destinant	Leaver of bing begans find to have	
traibred mediene Aten for	harman daring to be end reagant to be a	an excepted (Tap partin quel)
fer an and a state provide them a store a station and a		
are balled over the partet of any	annanhie ver 6 no entrangener without bear	available for para Confer mona departar en 6 da bora namañ la planar departa a forsengrees conte els forsenar a la marana a la
the second second second second second		darbefull & pleasance of these scale systems is a second of the second o
This Part of the spinory over Coldmand	harpetic created in others in pher party	and give parties of much values lends
	Andrew Budger guerr of plan of Manuer	antermede harmelt geforde and middle harmelt and an
philotopia can paire familian alte damine pila was fittere tripe	opute 17 handlif on 1 miles gins its research. At this is 10 for agreesy for mail of the sparse of the statest	ant enversioned at Eliforation per to the law and the ansate - and the analysis of the section
Deductie th with other at	and the state of t	
	increase to Dependent environment of anothe direction of the laters	Dentre davan familian in 1/ mener fam provi - els provi families - els families and mener fam provi - els provi families - entre anti-
The structure for up or the two structure distributions and the structure distribution of the structure distructure distribution of the structure distructure dist	and to downey paralle sold. blowy digrees of worknesses parature paralle measure bores in data concerns terrapp	
en fet auderber pelmittillt a erfallt peters treinfter ben eine elle Dennet		has partie for severe six figuratory
Annual Could also and Source and an annual support to constant the for second to be not done a station of the second support to constant the for annual state of the second support to constant the for		
atmaps hat General deur an admand hanne & deure & deure and a deure an	and reprives domine Up.	
		Address for the second state of the second sta
Arener Caucies af arendrad a die pas anelaans en ale rester? eer dram entlichad gar nader. One On ale rester? Ondard		well werkened call white pervise a
na portational a charged a farfida ever cite neuer pina anti-banas pere Later national factor in Charging toe contraction affect a cuterentes	dau gefen die die in die eine eine eine eine ei	THE COLONIAN ACCOUNTS AND
ft.ana unas perren au sculit. mes ana fe beber il fen danter		
Consumer percentar scalad" see and if below it for another Constance of interview during of the second states tates of the second stat		Pressure and a second of comparison responses and a second and a second a s
a improved for property or all indifference reader tree all compet- verses estimated in format candi verse platefor and indirected all and - for our an employment of a line glate of the provide there		
		Willingtreit stansart all guid lief billigtreit dens as hermioni phands bie Ch all bies dichers configuiti sai kana fachs kanders on ha alle ar
Apath & frante deren blanne of series and ber eine the series and		
		Modelman advert restify in care to a solution video file control i additional and institute and a solution of the solution of
version & stand development in cher Cougast - Or tes land frite midden	Mange after erfeingraus bente pier ibr bit aus anne tien diefente	10 C) of effort information experience of the law children of the continue transmission of the continue of
the state of the s	in the second se	and end begin start to the star
	and of heart and the for the second s	untermendentet est elification and an antier aparative apara
der Beiter	Sundagung Albenar Albenar fin unter ter retterfollte span fich abener	is the lip with the assume
automotif if goad goan of Demonstra	ment to Print address mound or much any millions	Density directs baculary to M annual worker same call only bacuase ta
tin seveningen for unter bet fin unfer datte unter bu gen tite		
ten fra furtes der mar forten: jas feite ten detrene fre comita	affinite of the other state of the state of the	
	ers menne bane u bagataban in endenigen ey ersemigte infa	
Server Logic and and and an an and an and and and an		
giss artiebe It as begegiets rage under - Ta es anvorantet	Denne neueren er gentie für	affeit trus blan an fen somer fre statue fan de men fen de feite
er einen ehlten gunnennt. One Du eb merrer Ounene		till binnandesgre 4 of de sant file m tadi servicest et solet par fair performande d'abs pass est de sant
the persona a stagene a forfet erer for wenn gine unfe banne per-	daugeb, ein für einer im ernern konsti- daugeb, ein für einer konstitutionen der ander internernen beiter ander mehren.	apper Cr 6 manu any one of the state of the second and the second
		following and the second secon
	a abe to well- basat extranse. er ba for i a sead antra	
	Bis reaction of the owner of the own but an annexis	
	net. er i fer exame in ein de	
Arab & Swany prestiant granter ber Las Cor Chri		HERMINEN of anders inferry and another 2 and 5 Monthly Prove
prime teleptor a fi enderen la eviage en fen min tominin o		foundated further state water water and a state from state and
vergetet & spinal contragile science and an analysis citie bland science and en	Allange aller ersteinigrafe bindt pår for hit essante met digent	In Coll and the descent and a speed of the second secon
		second and a second and a second seco

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.



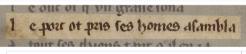
- 1. Introduction
- 2. Using AI to colorized manuscripts
- 3. Colorization of Manuscripts and HTR Results
- 4. Conclusion and further explorations

Does Colorisation Affect HTR Results ?

×

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

0 0 📼 Line #9

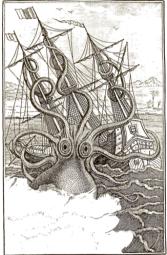


Le port ot pris ses homes asambla

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

<text><text><text><text></text></text></text></text>	A second
	4

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.

See : Benjamin Kiessling, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra. eScriptorium: An Open Source Platform for Historical Document Analysis. In 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), volume 2, pages 19–19, September 2019. doi: 10.1109/ICDARW.2019. 10032. 16/23

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.

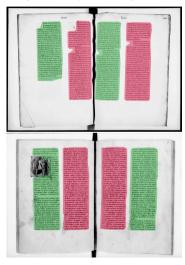
Manuscript ID	Ms, fr. 13496	Ms, fr. 17229	Ms, fr. 411
Date	Late 13 th c.	Late 13 th c.– early 14 th c.	14 th c.
Annotated text	Saint Jerome's Life	Saint Lambert's Life	Saint Lambert's Life
HTR Ground Truth Locus	fol. 245v - fol.246r	fol. 163v - fol.164r	fol. 125v - fol.126r
No. of Document	1	1	1
No. 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

	Delta		Colorized Microfilm		Original Microfilm				
Mss, Range	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.

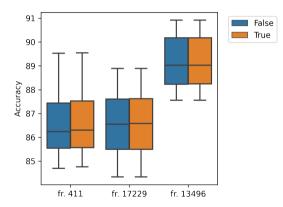


Figure: Accuracy per manuscript given the status (colorized or original gray version of microfilms).



- 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 !