Multiple Document Datasets Pre-training Improves Text Line Detection With Deep Neural Networks

Mélodie Boillet[1,2], Christopher Kermorvant[1,2] and Thierry Paquet[2]

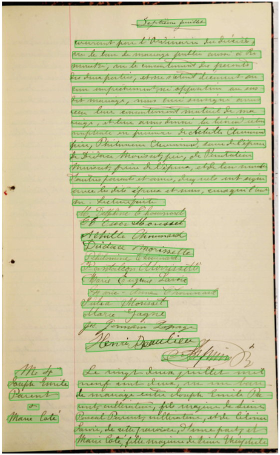
[1]Teklia SAS, Paris, France
[2]LITIS, Rouen-Normandy University, France

ICPR - 12th January 2021

Presentation overview

Text line segmentation

- **Goal:** detect the text lines of an image;
- Application: apply a text recognition system on the detected text lines.



Problems of state-of-the-art system dhSegment

- Needs a lot of annotated data;
- Good results but can still be improved;
- Too long to analyse a whole corpus:
  $\sim$ 66 days for 2M images (on a GPU GeForce RTX 2070 8G for Balsac corpus).

Is pre-training on natural scene images the most suitable for working on document images?

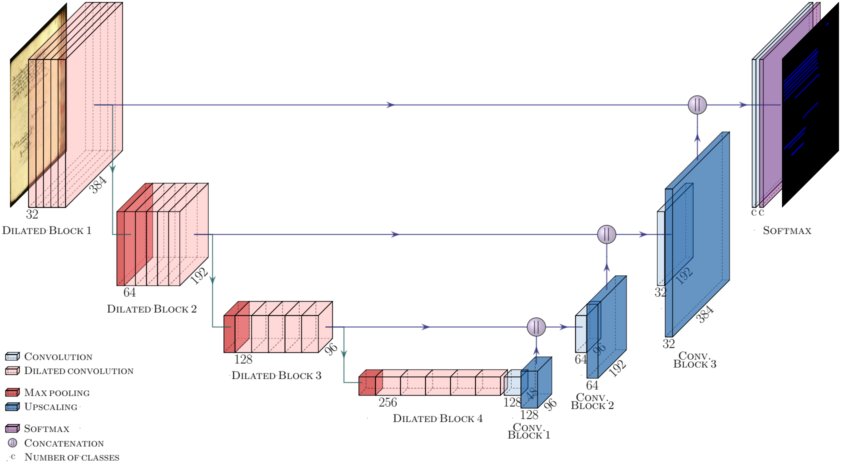
Main goal

Analyse the impact of a pre-training step on the line segmentation task.

We want a model:

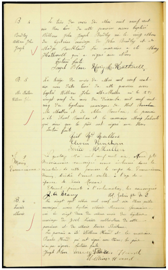
- Containing no pre-trained part learnt on natural scene images;
- Having less parameters than SOTA on historical documents (dhSegment) and a reduced prediction time;
- Yielding higher accuracy than SOTA on historical documents (dhSegment).

Architecture of our Doc-UFCN - inspired by [Yang2017]



Datasets

Balsac: 913 annotated images

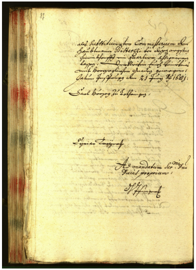


Pages of acts extracted from Quebecois registers.

Horae: 557 annotated images [Boillet2019]

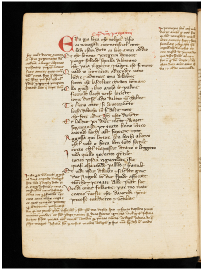
Pages extracted from 500 digitized books of hours.

READ-BAD: 2036 annotated images [Grüning2017]



Archival documents written between 1470 and 1930.

DIVA-HisDB: 120 annotated images [Simistira2016]



Handwritten pages extracted from 3 medieval manuscripts.

Comparison with dhSegment

Dataset	Model	IoU	Pr	Rec	F1	Time[1]
Balsac	dhSegment	73.78	92.07	78.76	84.81	66.3
	Doc-UFCN	83.79	94.80	87.86	91.11	9.2
Horae	dhSegment	65.22	71.70	89.29	82.32	18.8
	Doc-UFCN	63.95	78.38	80.45	84.93	2.3
READ-Simple	dhSegment	64.55	85.04	71.85	77.25	8.4[2]
	Doc-UFCN	64.03	81.76	75.60	76.66	1.0[2]
READ-Complex	dhSegment	52.91	79.28	59.16	69.27	10.6[2]
	Doc-UFCN	54.40	83.62	61.97	73.16	1.3[2]
DIVA-HisDB	dhSegment	74.24	92.41	79.10	85.19	N/A
	Doc-UFCN	75.71	92.14	80.88	86.09	N/A

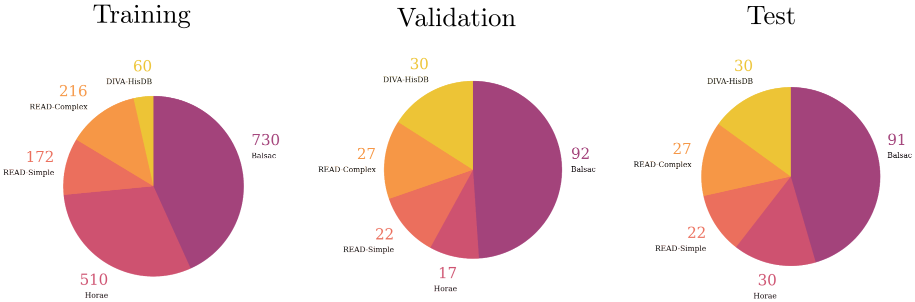
	dhSegment	Doc-UFCN
Number of parameters	32.8M(9.36M)	4.1M

[1] Prediction time (GPU GeForce RTX 2070 8G) in days to analyse the whole corpus.

[2] Estimation based on the manuscripts sizes without *BHIC* and *Unibas*.

Split of the *Multiple document dataset*

Does pre-training on document images improve the performances?



Comparison with dhSegment: impact of pre-training

Data	Model	IoU	Pr	Rec	F1
Balsac	dhSegment	73.78	92.07	78.76	84.81
	dhSegment PT	74.02	91.89	79.09	84.95
	Doc-UFCN	83.79	94.80	87.86	91.11
	Doc-UFCN PT	84.87	94.25	89.49	91.75
Horae	dhSegment	65.22	71.70	89.29	82.32
	dhSegment PT	60.69	80.94	73.65	81.99
	Doc-UFCN	63.95	78.38	80.45	84.93
	Doc-UFCN PT	68.81	80.31	84.80	88.62
READ-Simple	dhSegment	64.55	85.04	71.85	77.25
	dhSegment PT	65.07	88.34	71.56	80.72
	Doc-UFCN	64.03	81.76	75.60	76.66
	Doc-UFCN PT	68.14	83.19	78.05	79.45
READ-Complex	dhSegment	52.91	79.28	59.16	69.27
	dhSegment PT	53.34	85.51	57.80	68.47
	Doc-UFCN	54.40	83.62	61.97	73.16
	Doc-UFCN PT	60.28	81.03	68.17	78.30
DIVA-HisDB	dhSegment	74.24	92.41	79.10	85.19
	dhSegment PT	73.00	91.56	78.28	84.32
	Doc-UFCN	75.71	92.14	80.88	86.09
	Doc-UFCN PT	74.72	89.43	82.20	85.44

Conclusion

Does pre-training on document images improve the performances?
YES

Intersection-over-Union:

- ✓ +5 percentage points on Horae and READ-Complex;
- ✓ +4 percentage points on READ-Simple;
- ≈ Similar performances on Balsac;
- ✗ −1 percentage point on DIVA-HisDB.

Our results are overall better than dhSegment ones (except for the precision metric).

Conclusion

We designed a model:

- Lighter than dhSegment;
- Giving on average better results;
- Having a reduced prediction time: up to 8 times faster.

\+ We have shown that pre-training on various historical documents can improve the performances.

Future work:

- Test our architecture on other tasks than text line detection;
- Build an historical model trained on a large dataset of diverse historical documents.

Conclusion

Thanks for your attention!

Questions?

Bibliography

[Oliveira2018] Sofia Ares Oliveira, Benoit Seguin, and Frederic Kaplan. "dhSegment: A generic deep-learning approach for document segmentation". In: *Frontiers in Handwriting Recognition (ICFHR), 2018 16th International Conference on*. IEEE. 2018, pp. 7–12

[Yang2017] Xiao Yang et al. "Learning to Extract Semantic Structure from Documents Using Multimodal Fully Convolutional Neural Network". In: vol. abs/1706.02337. 2017. arXiv: 1706.02337. URL: http://arxiv.org/abs/1706.02337

[Boillet2019] Mélodie Boillet et al. "HORAE: An Annotated Dataset of Books of Hours". In: *Proceedings of the 5th International Workshop on Historical Document Imaging and Processing*. HIP '19. Sydney, NSW, Australia: Association for Computing Machinery, 2019, 7–12. ISBN: 9781450376686. DOI: 10.1145/3352631.3352633. URL: https://doi.org/10.1145/3352631.3352633

[Grüning2017] Tobias Grüning et al. "READ-BAD: A New Dataset and Evaluation Scheme for Baseline Detection in Archival Documents". In: *CoRR* abs/1705.03311 (2017). arXiv: 1705.03311. URL: http://arxiv.org/abs/1705.03311

[Simistira2016] F. Simistira et al. "DIVA-HisDB: A Precisely Annotated Large Dataset of Challenging Medieval Manuscripts". In: *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. Oct. 2016, pp. 471–476. DOI: 10.1109/ICFHR.2016.0093