Towards End-to-end Handwritten Document Recognition

Thesis defense, 2022/09/29

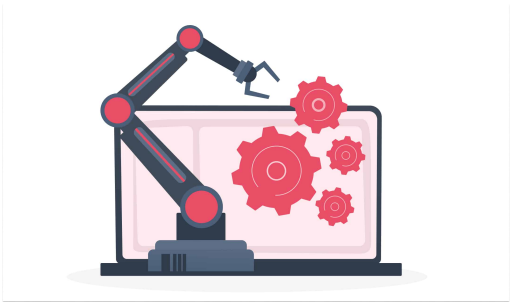
Denis Coquenet

LITIS - EA 4108 University of Rouen, France

Rapporteurs :	Christian Wolf	MCF/HDR (LIRIS, Lyon)
	Mathieu Aubry	MCF/HDR (LIGM, Paris)
Examinateurs :	Harold Mouchere	Professeur (LS2N, Nantes)
	Elisa Fromont	Professeure (IRISA/INRIA, Rennes)
Invitée :	Nihel Kooli	Représentante DGA
Directeur de thèse :	Thierry Paquet	Professeur (LITIS, Rouen)
Co-encadrant :	Clément Chatelain	MCF/HDR (LITIS, Rouen)

Table of contents

Automation age



Industry

Finance

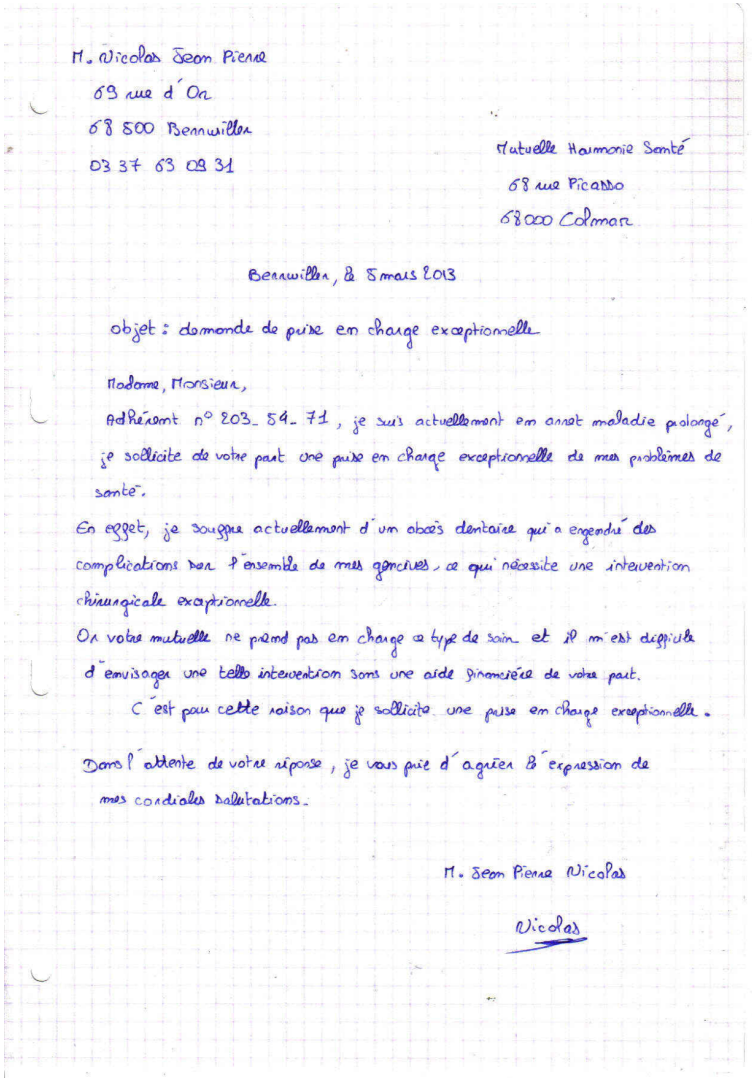
Transport

Automatic understanding of documents: a need for many applications

Examples

- Industry: automation of the analysis of bank checks, forms, invoices
- Academic: digitization of student's handwriting
- Humanities: transcription of historical documents
- Military: real-time document translation

Understanding documents: a complex task



M. Nicolas Jean Pierre
69 rue d'Or
68 800 Bennwiller
03 37 63 09 31

Mutuelle Harmonie Santé
68 rue Picasso
68000 Colmar

Bennwiller, le 5 mars 2013

objet : demande de prise en charge exceptionnelle

Madame, Monsieur,

Adhérent n° 203_59_71, je suis actuellement en arrêt maladie prolongé, je sollicite de votre part une prise en charge exceptionnelle de mes problèmes de santé.

En effet, je souffre actuellement d'un abcès dentaire qui a engendré des complications sur l'ensemble de mes gencives, ce qui nécessite une intervention chirurgicale exceptionnelle.

Or votre mutuelle ne prend pas en charge ce type de soin et il m'est difficile d'envisager une telle intervention sans une aide financière de votre part.

C'est pour cette raison que je sollicite une prise en charge exceptionnelle.

Dans l'attente de votre réponse, je vous prie d'agréer l'expression de mes cordiales salutations.

M. Jean Pierre Nicolas

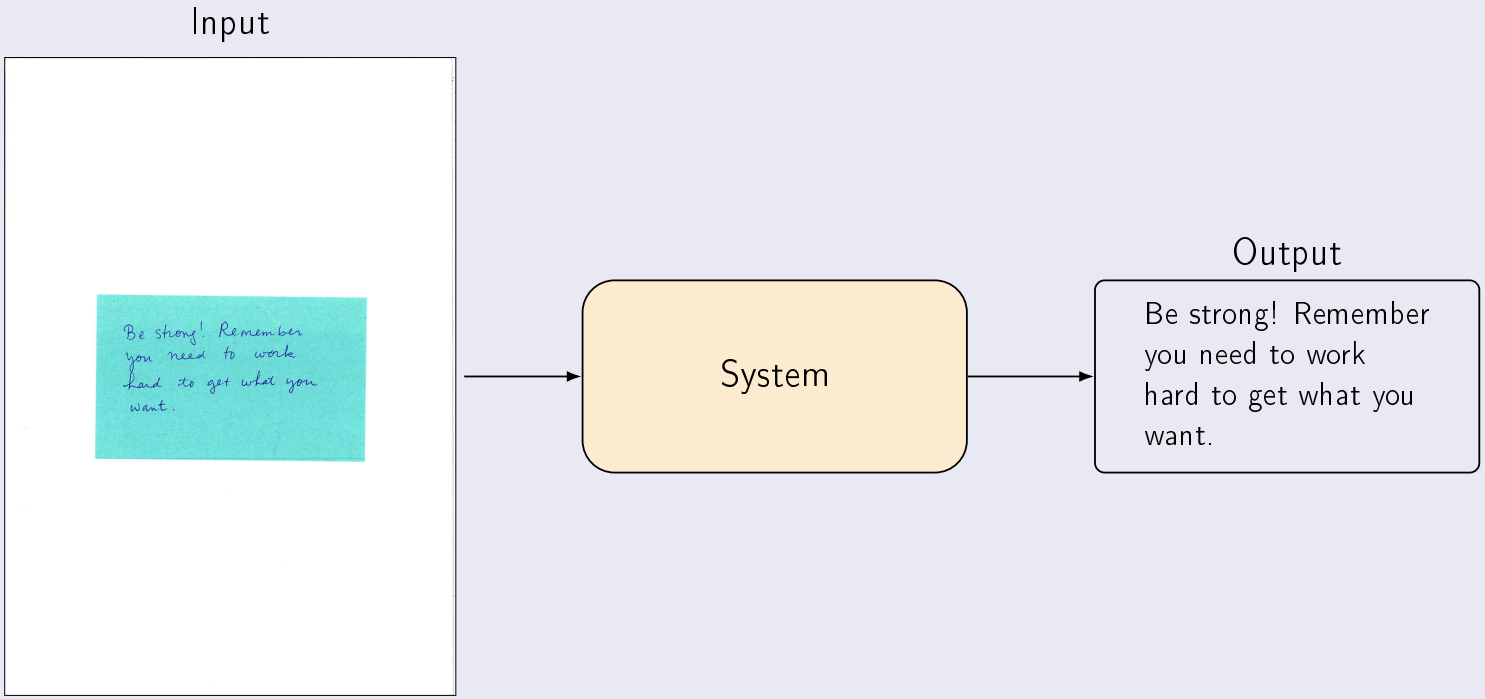
Nicolas

- Type: letter
- Language: French
- Structure: layout, reading order
- Content: text, table, image, graph
- Semantic: location, date, signature

Image from the MAURDOR dataset [1]

Handwritten Text Recognition (HTR)

An image-to-sequence problem



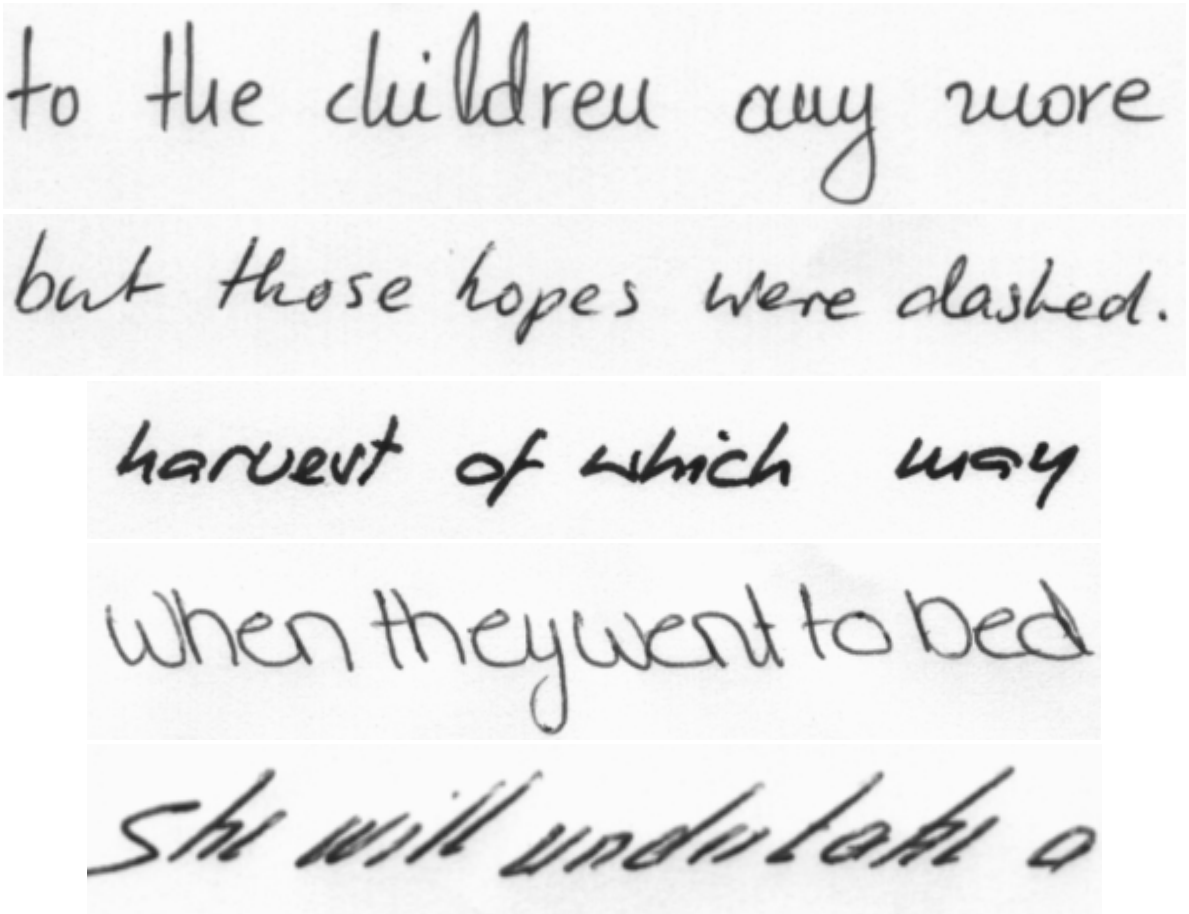
Input: an image

Output: a sequence of characters

Challenges

A wide variety of documents

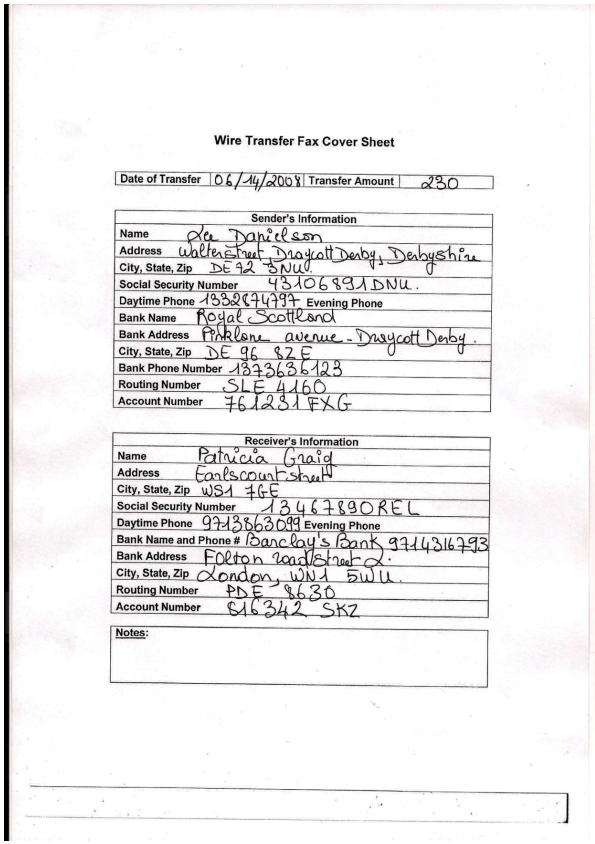
Writing styles, layout, size / resolution, background



Challenges

A wide variety of documents

Writing styles, **layout**, size / resolution, background



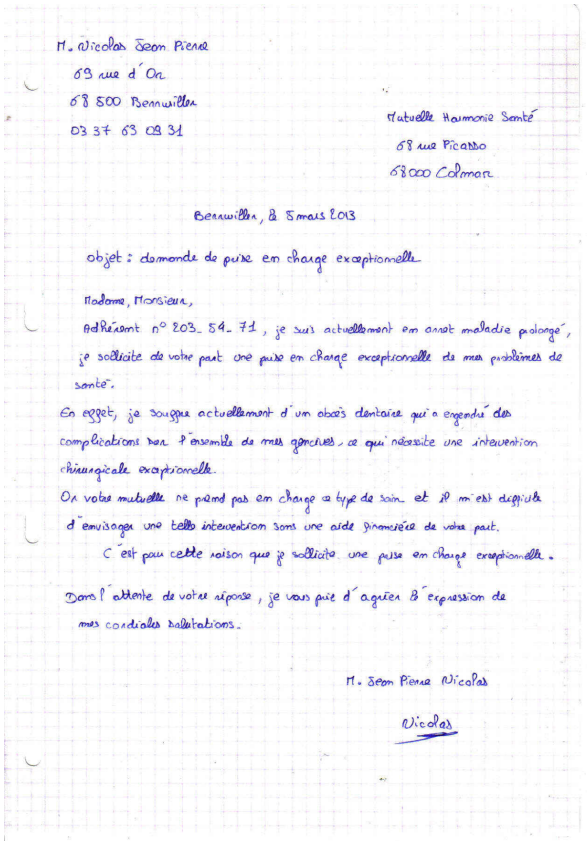
Wire Transfer Fax Cover Sheet

Date of Transfer	06/14/2008	Transfer Amount	280

Sender's Information	
Name	Rex Danielson
Address	Walter Street Draycott Derby, Derbyshire
City, State, Zip	DE 92 5NU
Social Security Number	43106891DNU.
Daytime Phone	1332674797 Evening Phone
Bank Name	Royal Scotland
Bank Address	Pinklone avenue - Draycott Derby.
City, State, Zip	DE 96 8ZE
Bank Phone Number	1373636123
Routing Number	SLE 4160
Account Number	761231 FXG

Receiver's Information	
Name	Patricia Graid
Address	Earlscourt Street
City, State, Zip	WS1 7GE
Social Security Number	1346789OREL
Daytime Phone	9713863099 Evening Phone
Bank Name and Phone #	Barclay's Bank 97143167 93
Bank Address	Folton road Street 2.
City, State, Zip	London, WN1 5WU.
Routing Number	PDE 8630
Account Number	816342 SKZ

Notes:



M. Nicolas Jean Pierre
69 rue d'Or
68 500 Bennwillen
03 37 63 09 31

Mutuelle Harmonie Santé
68 rue Picasso
68000 Colmar

Bennwillen, le 5 mars 2013

objet : demande de prise en charge exceptionnelle

Madame, Monsieur,

Adhérent n° 203_54_71, je suis actuellement en arrêt maladie prolongé, je sollicite de votre part une prise en charge exceptionnelle de mes problèmes de santé.

En effet, je souffre actuellement d'un abcès dentaire qui a engendré des complications sur l'ensemble de mes gencives, ce qui nécessite une intervention chirurgicale exceptionnelle.

Or votre mutuelle ne prend pas en charge ce type de soin et il m'est difficile d'envisager une telle intervention sans une aide financière de votre part.

C'est pour cette raison que je sollicite une prise en charge exceptionnelle.

Dans l'attente de votre réponse, je vous prie d'agréer l'expression de mes cordiales salutations.

M. Jean Pierre Nicolas

Nicolas

Challenges

A wide variety of documents

Writing styles, layout, size / resolution, **background**

Challenges

A wide variety of documents

Writing styles, layout, size / resolution, background

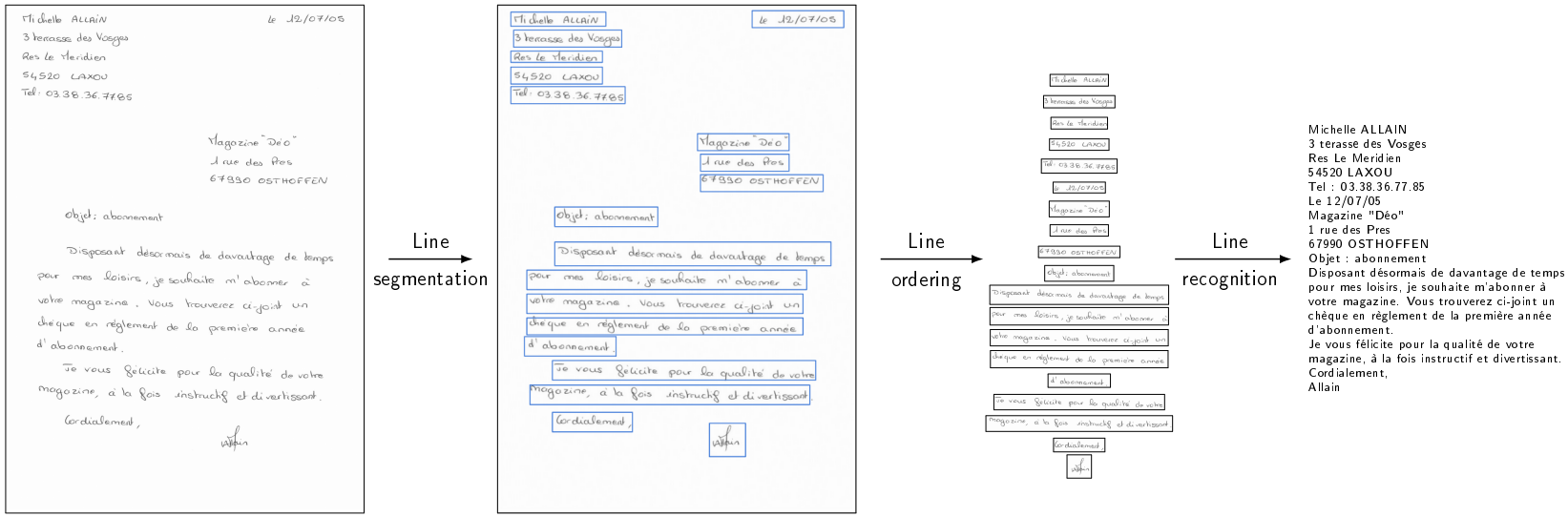
No a priori knowledge about the document

- Number of lines
- Number of characters per line
- Reading order

Table of contents

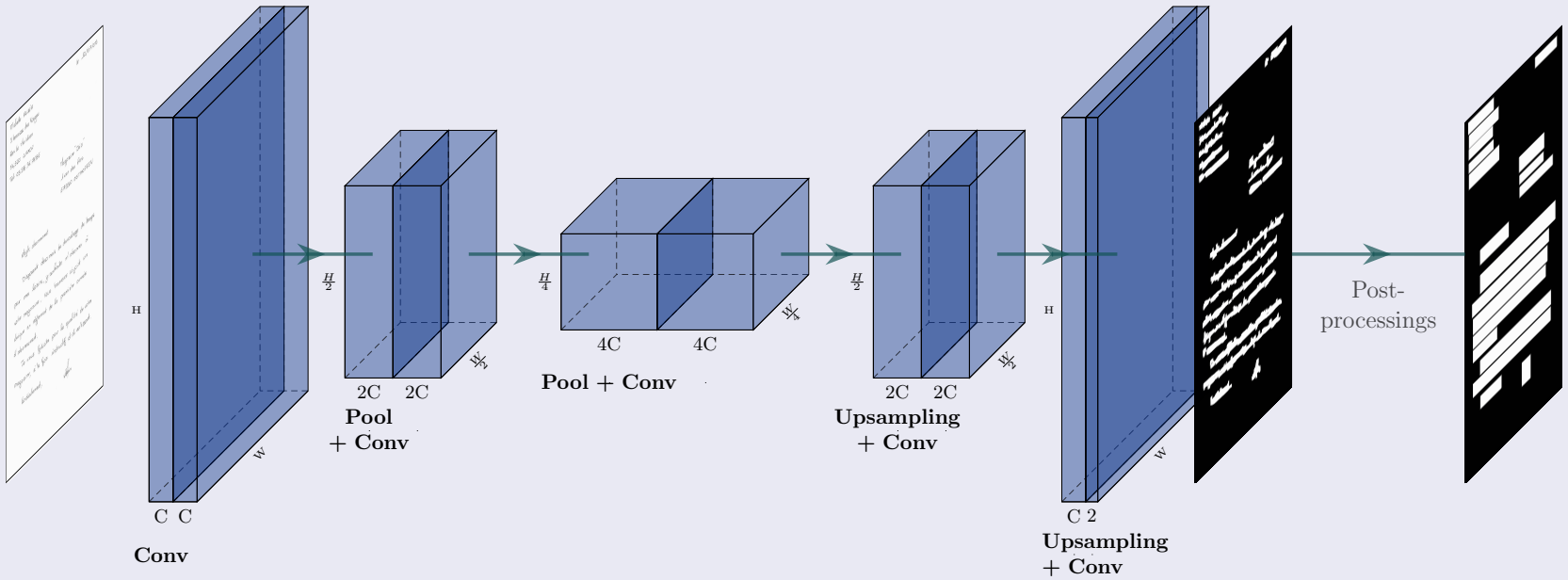
The line-level sequential paradigm

- Segmentation
- Ordering
- Recognition



Related works: Segmentation stage

Text line segmentation architecture (FCN) [2, 3]

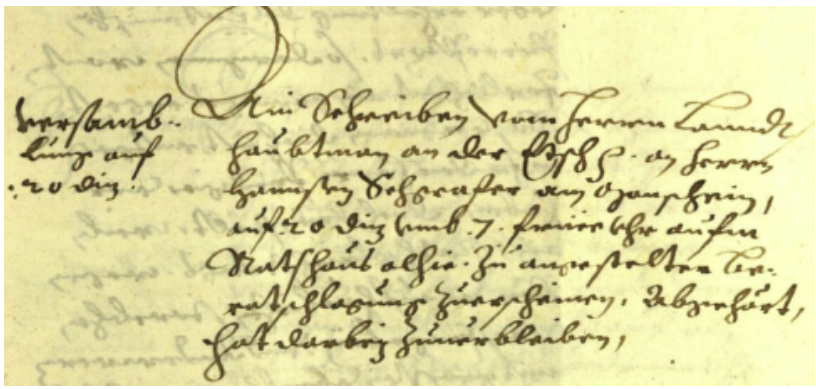


[2] Renton *et al.*, IJDAR 2018
[3] Boillet *et al.*, IJDAR 2022

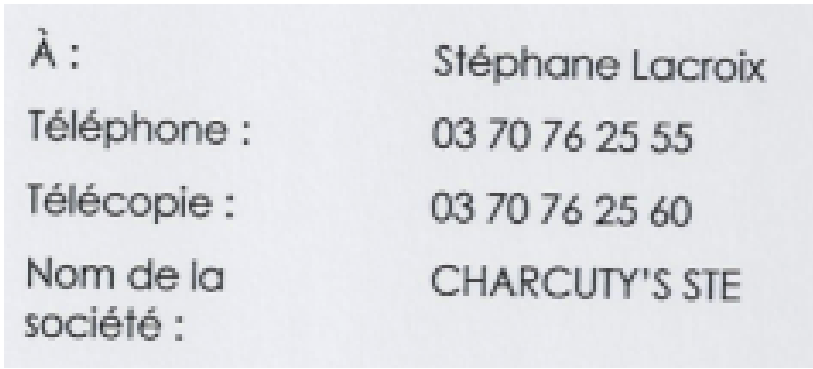
Related works: Ordering stage

A rule-based approach

From top to bottom and from left to right for most Latin languages.



(a) Expected reading order by column.

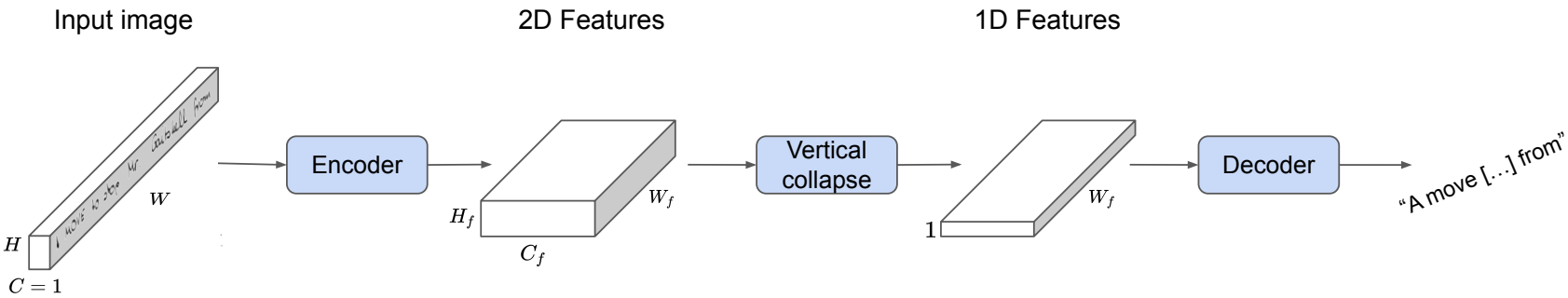


(b) Expected reading order by row.

Related works: Recognition stage

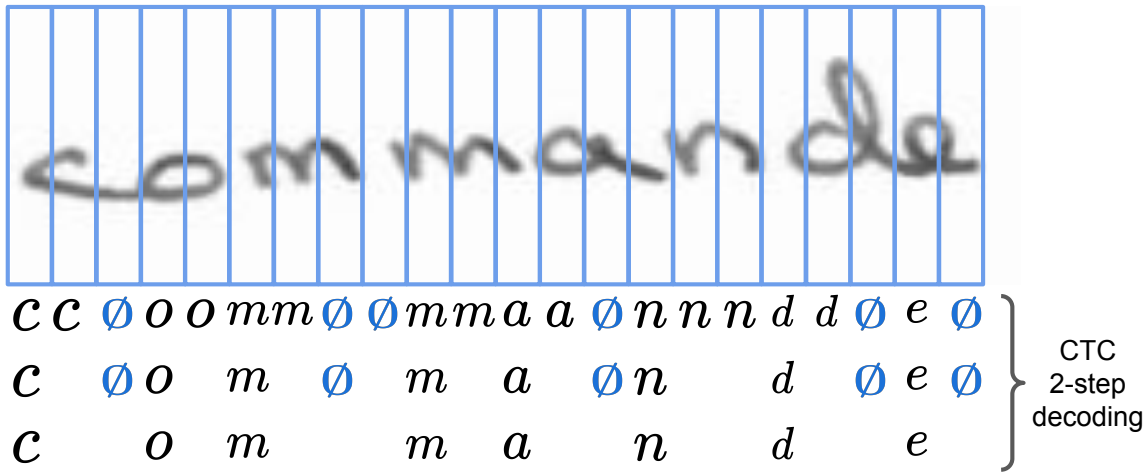
Challenges:

- going from a 2D input image to a 1D sequence of characters
- a variable, unknown number of ordered characters to predict



Related works: Recognition stage

The Connectionist Temporal Classification (CTC) paradigm [4]



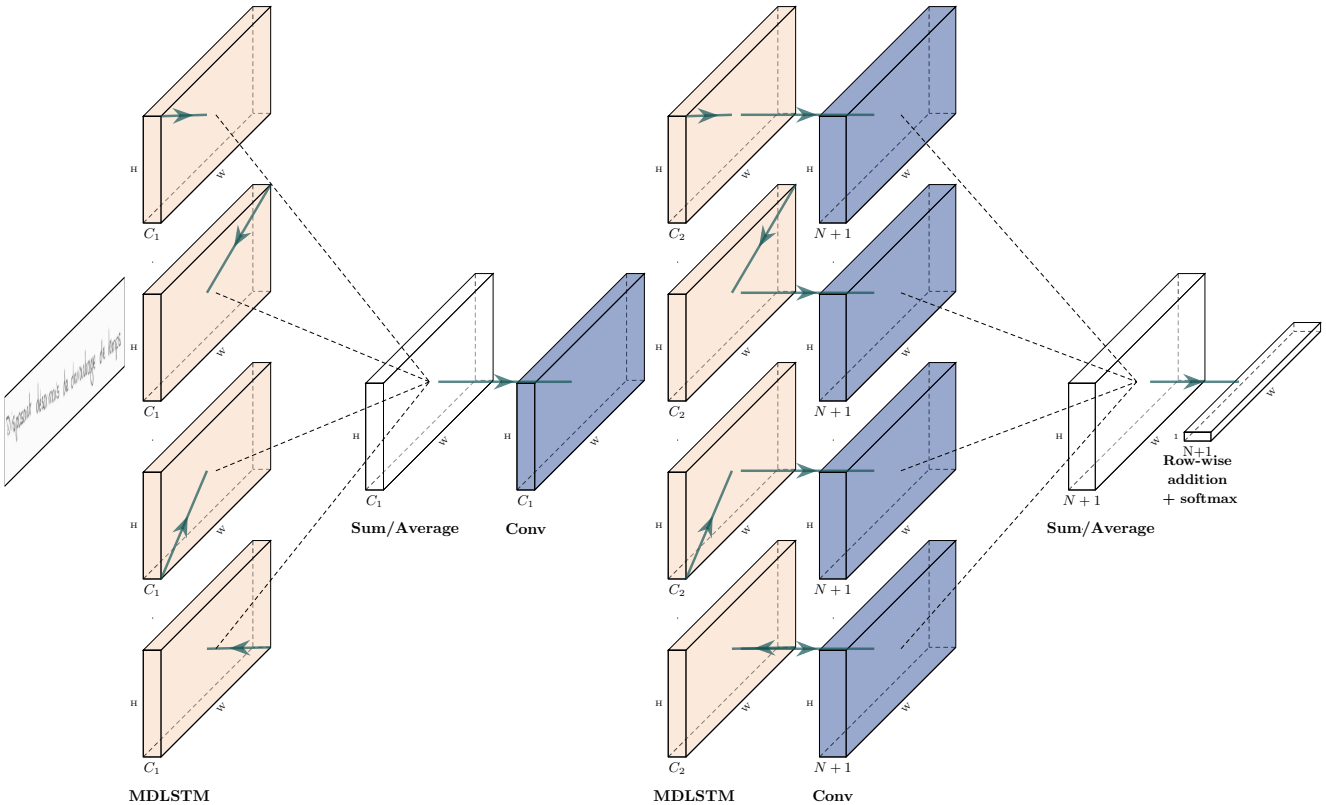
- A frame-by-frame decision process
- Special blank token $\varnothing$
- A left-to-right constrained alignment

➤ CTC loss

[4] Graves *et al.*, ICML 2006

Related works: Recognition stage

Architectures: MDLSTM [5], **CNN+MDLSTM** [6, 7], CNN+BLSTM [8, 9], CNN [10, 11], FCN [12, 13]

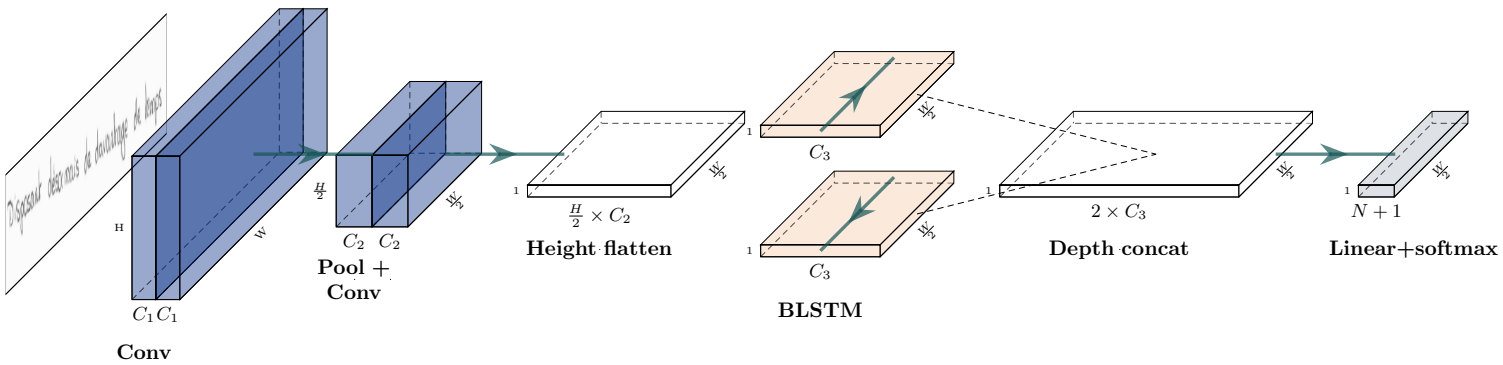


Convolutional Neural Network + Multi-Dimensional Long-Short Term Memory

[5] Graves *et al.*, NIPS 2008 [6] Pham *et al.*, ICFHR 2014 [7] Voigtlaender *et al.*, ICFHR 2016

Related works: Recognition stage

Architectures: MDLSTM [5], CNN+MDLSTM [6, 7], **CNN+BLSTM** [8, 9], CNN [10, 11], FCN [12, 13]

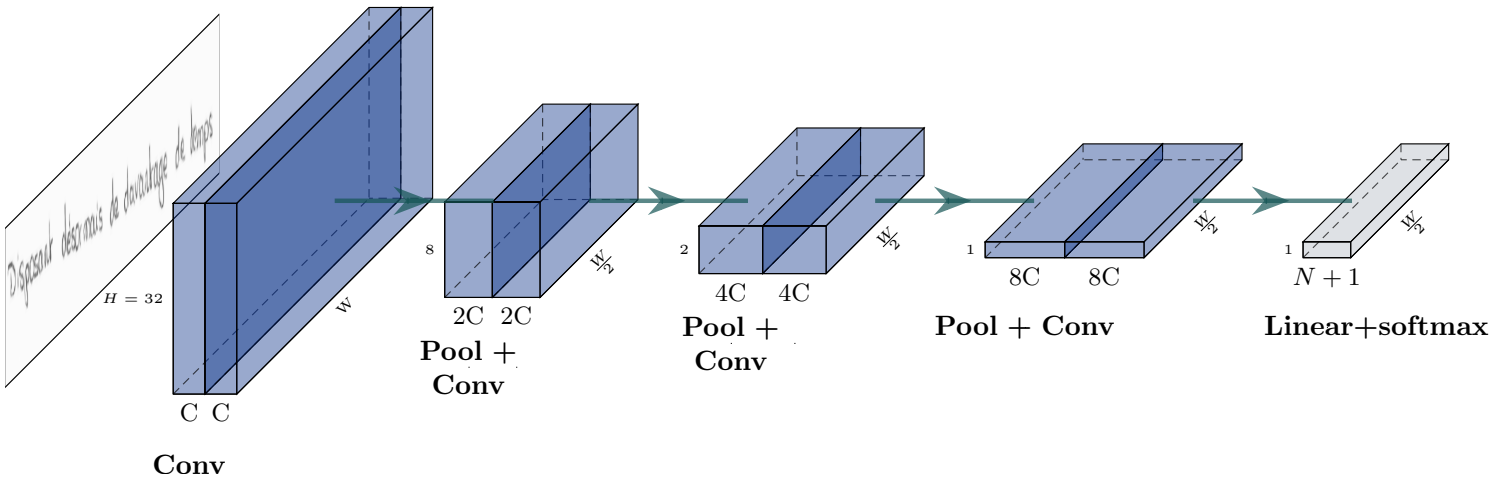


Convolutional Neural Network + Bi-directional Long-Short Term Memory

[8] Wigington *et al.*, ICDAR 2017
[9] Puigcerver *et al.*, ICDAR 2017

Related works: Recognition stage

Architectures: MDLSTM [5], CNN+MDLSTM [6, 7], CNN+BLSTM [8, 9], **CNN** [10, 11], FCN [12, 13]

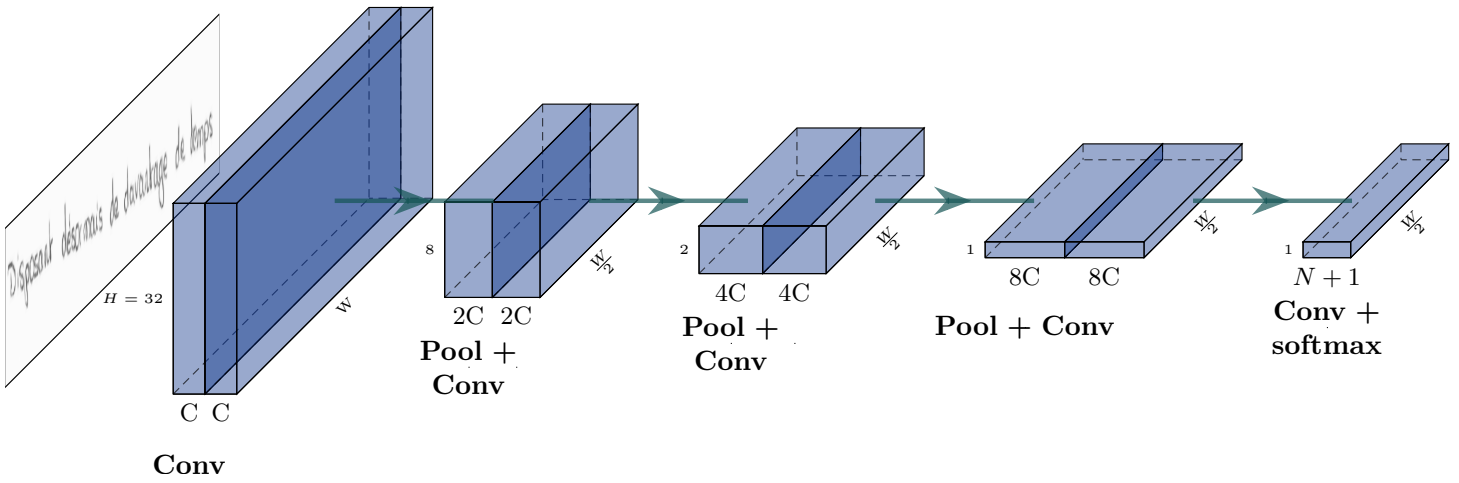


Convolutional Neural Network

[10] Ptucha *et al.*, PR 2019
[11] Coquenet *et al.*, WML@ICDAR 2019

Related works: Recognition stage

Architectures: MDLSTM [5], CNN+MDLSTM [6, 7], CNN+BLSTM [8, 9], CNN [10, 11], **FCN** [12, 13]

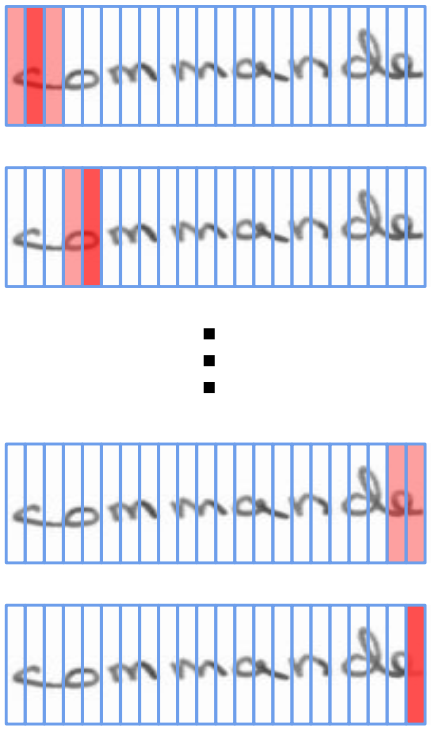


Fully Convolutional Network

[12] Yousef *et al.*, PR 2020
[13] Coquenet *et al.*, ICFHR 2020

Related works: Recognition stage

The attention paradigm (at character level) [14, 15]



- Iterative decoding process
- Implicit character segmentation
- Unconstrained attention
- Special end-of-transcription token <eot>
- ➤ Cross-Entropy loss

[14] Michael *et al.*, ICDAR 2019
[15] Wick *et al.*, ICDAR 2021

Conclusion

The sequential paradigm: a mature approach... with some limitations

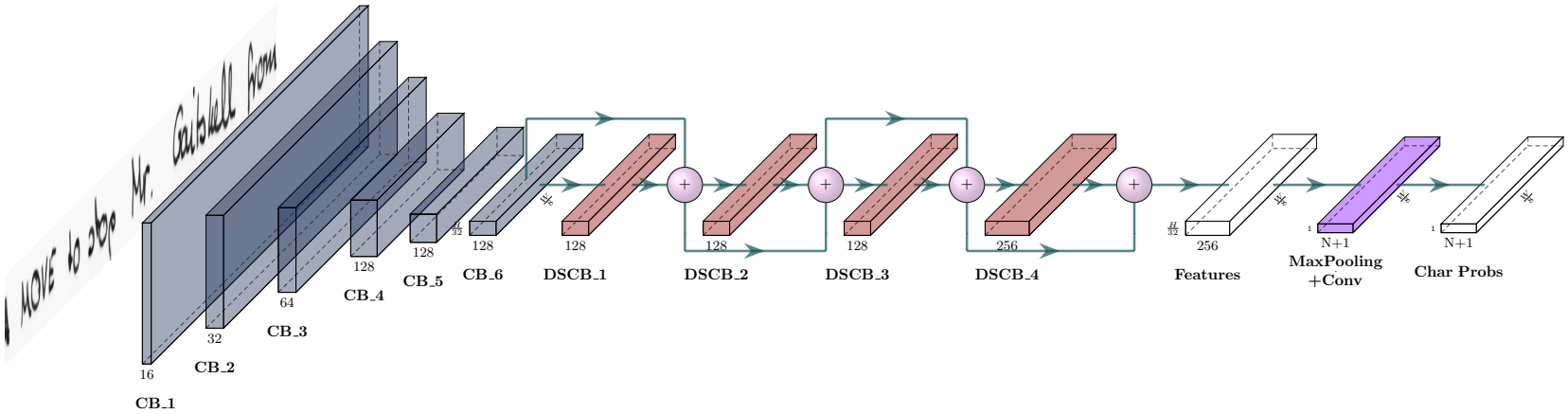
- Three steps treated independently
- A complex pipeline, hard to maintain
- Cumulative errors between steps
- Additional segmentation annotations
- Rule-based reading order

Goal: to overcome these limitations

Strategy: designing end-to-end HTR models step by step

- from line to document level

Contribution for line-level recognition (FCN) [19]



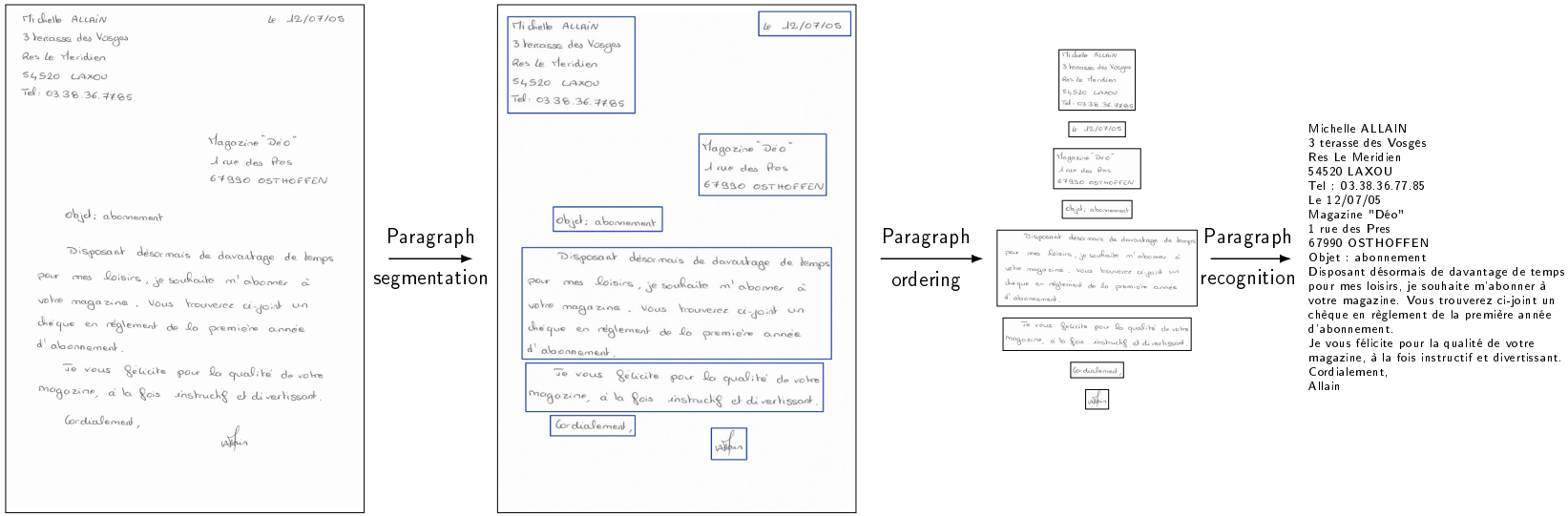
CB: Conv+Conv+Instance Norm.+Strided Conv
DSCB: DSC+DSC+Instance Norm.+DSC

Generic FCN encoder module for HTR

- Input of variable sizes
- Parallelizable operations
- Few parameters: 1.7 M
- Large receptive field: 961×337 px
- Competitive results on RIMES 2011 [16], IAM [17] and READ 2016 [18]

Table of contents

The paragraph-level sequential paradigm



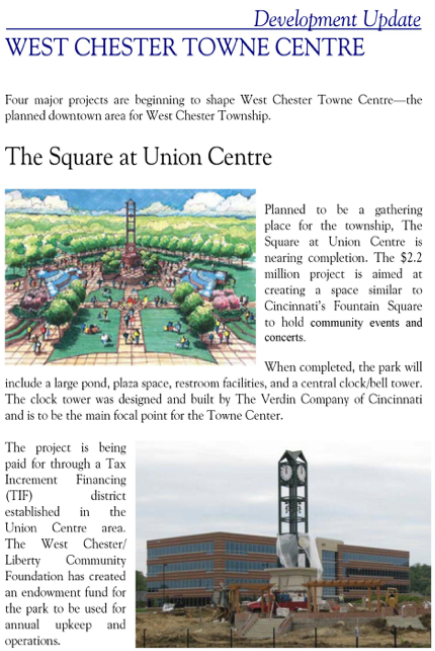
Goals

- Reduce segmentation annotations
- Benefit from larger context

Related works

Segmentation stage

Document Layout Analysis (FCN) [20, 21]



Development Update

WEST CHESTER TOWNE CENTRE

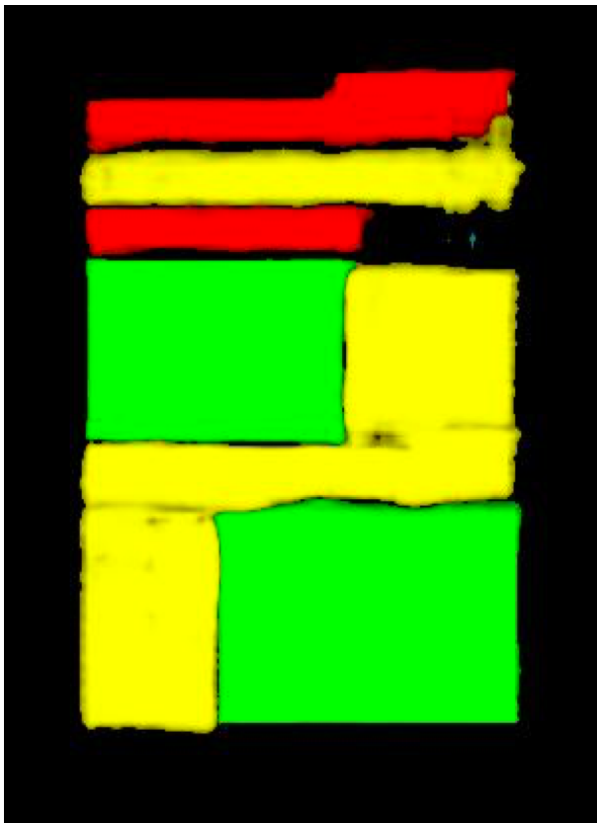
Four major projects are beginning to shape West Chester Towne Centre—the planned downtown area for West Chester Township.

The Square at Union Centre

Planned to be a gathering place for the township, The Square at Union Centre is nearing completion. The $2.2 million project is aimed at creating a space similar to Cincinnati's Fountain Square to hold community events and concerts.

When completed, the park will include a large pond, plaza space, restroom facilities, and a central clock/bell tower. The clock tower was designed and built by The Verdin Company of Cincinnati and is to be the main focal point for the Towne Center.

The project is being paid for through a Tax Increment Financing (TIF) district established in the Union Centre area. The West Chester/ Liberty Community Foundation has created an endowment fund for the park to be used for annual upkeep and operations.

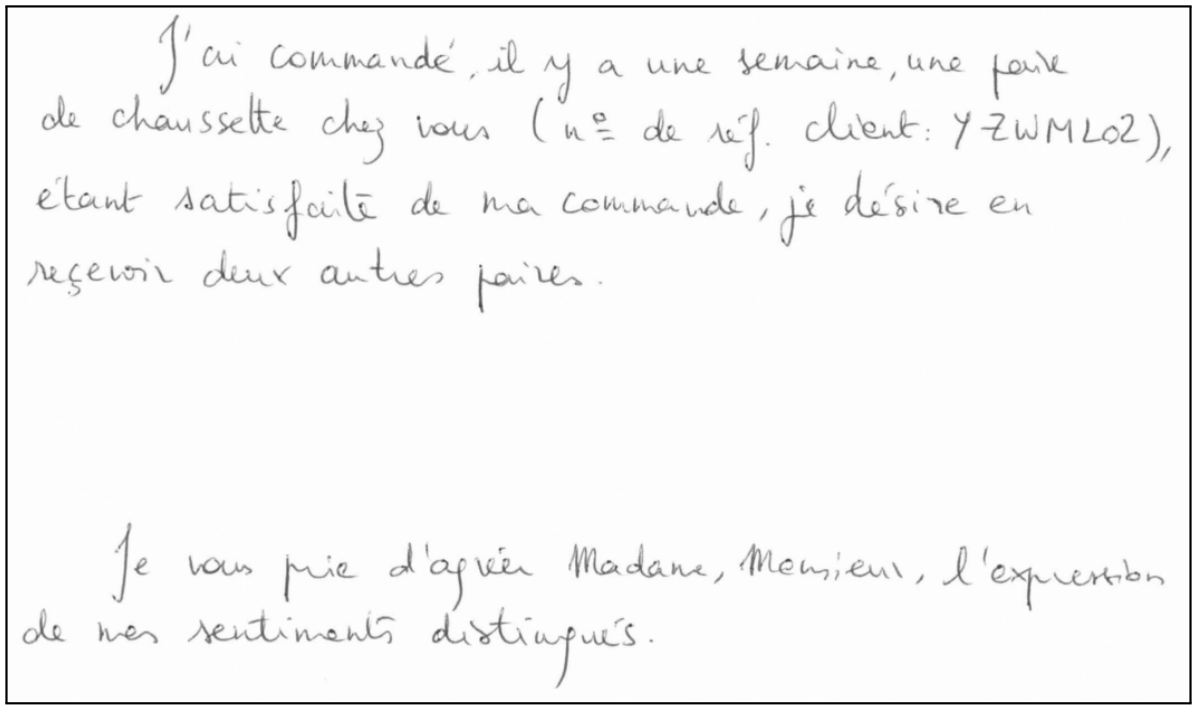


[20] Yang *et al.*, CVPR 2017
[21] Soullard *et al.*, PRL 2020

Related works: Paragraph recognition

Challenges from line to paragraph recognition

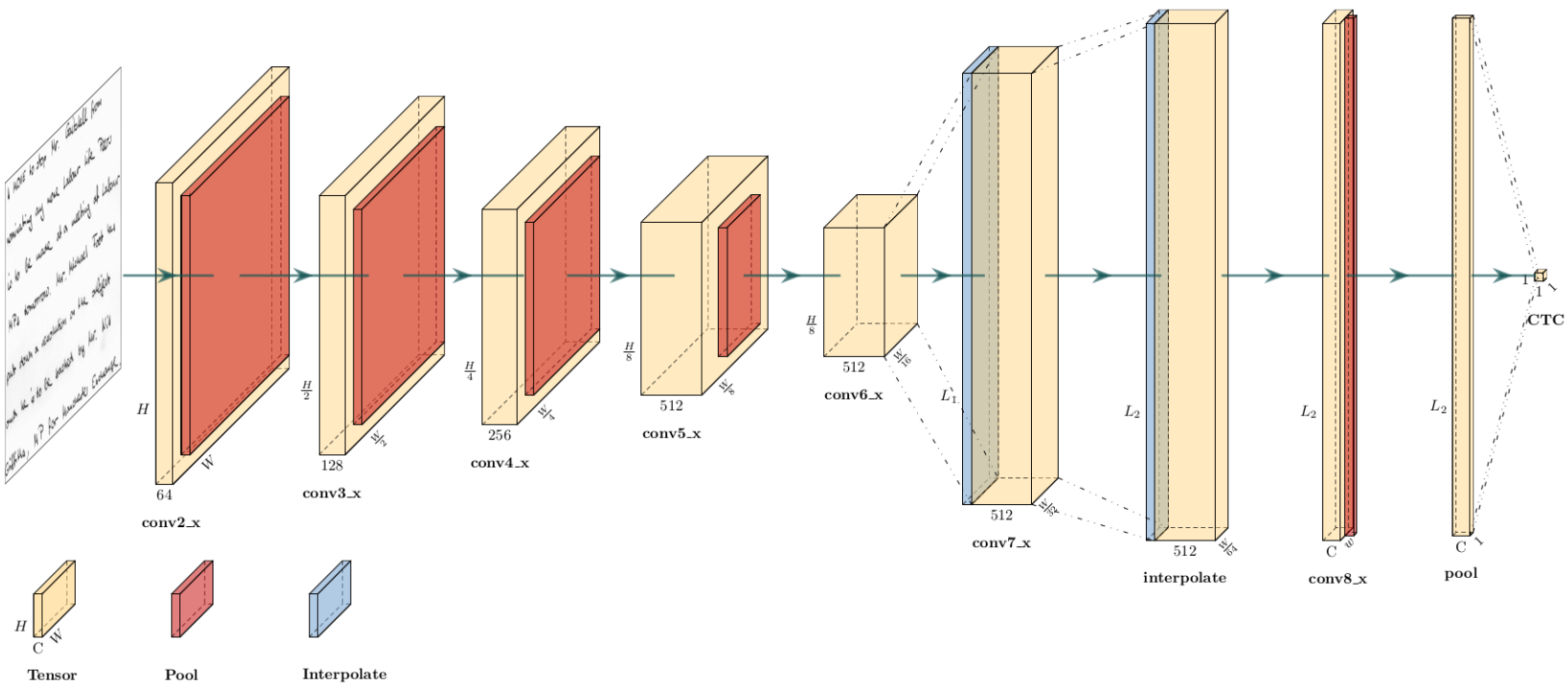
- An additional vertical reading order
- Variable number of text lines
- Variable interline spacing, indent



Related works: Paragraph recognition

CTC-only approaches

- OrigamiNet [22]

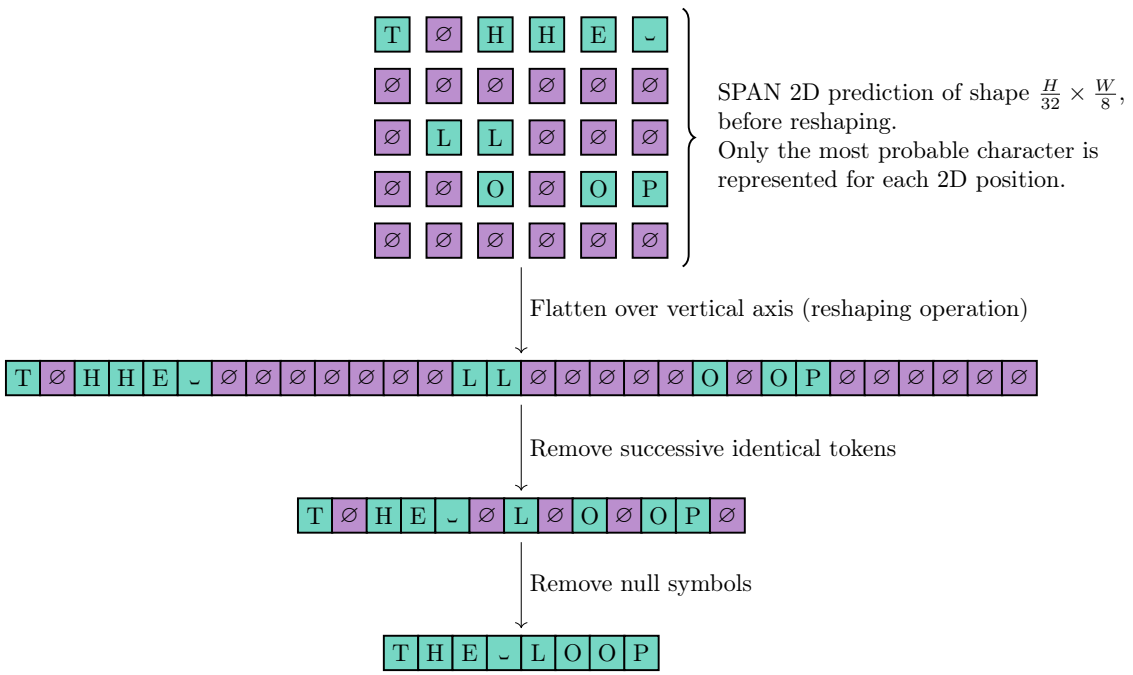


[22] Yousef *et al.*, CVPR 2020

Related works: Paragraph recognition

CTC-only approaches

- OrigamiNet [22]
- **Contribution:** Simple Predict & Align Network (SPAN) [23]



[23] Coquenet *et al.*, ICDAR 2021

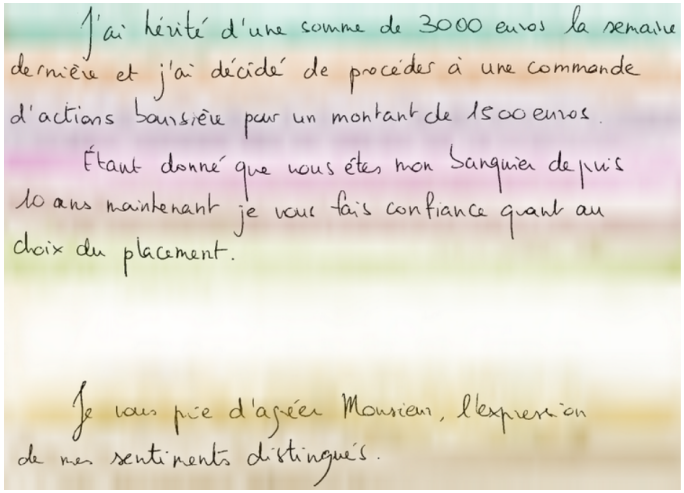
Related works: Paragraph recognition

CTC-only approaches

- OrigamiNet [22]
- **Contribution:** Simple Predict & Align Network (SPAN) [23]

Attention-based approaches

- Line-level attention [24]



[24] Bluche *et al.*, NIPS 2016

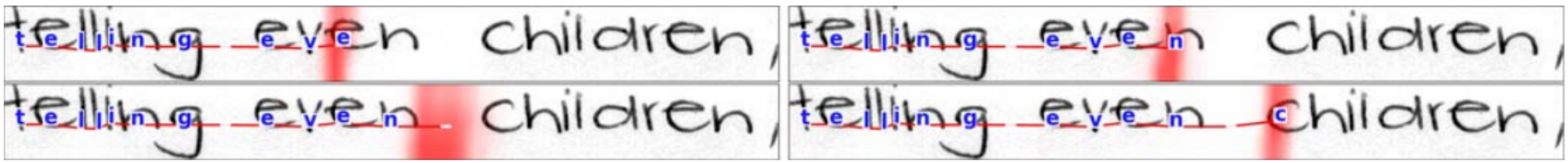
Related works: Paragraph recognition

CTC-only approaches

- OrigamiNet [22]
- **Contribution:** Simple Predict & Align Network (SPAN) [23]

Attention-based approaches

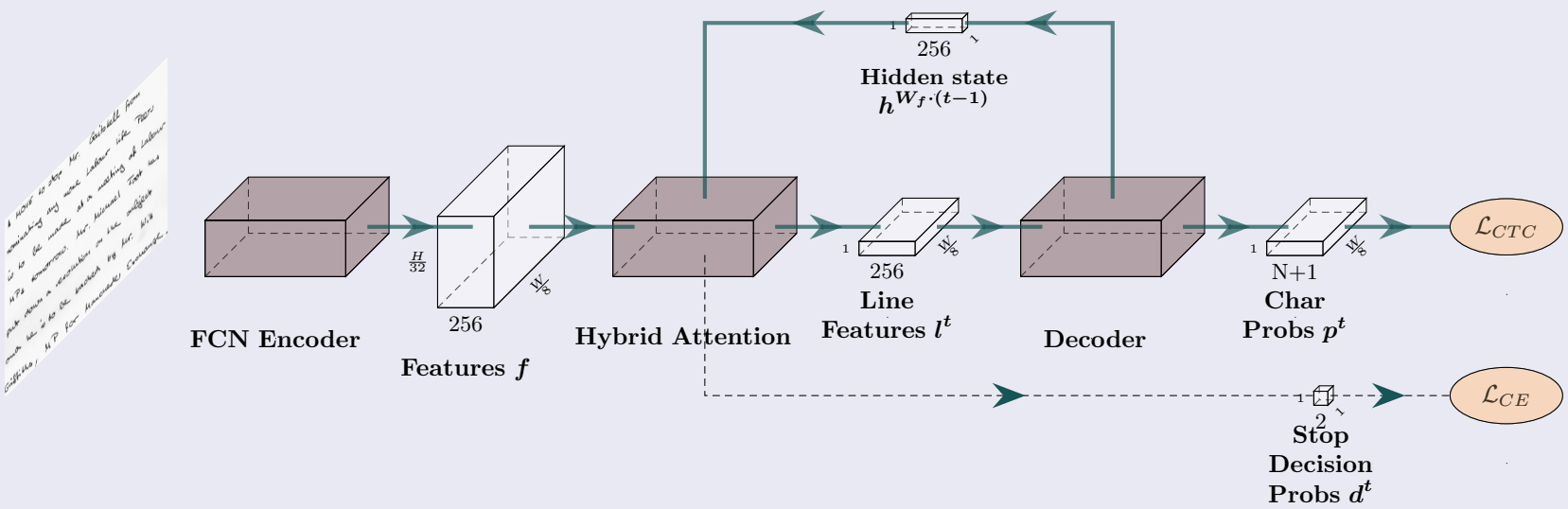
- Line-level attention [24]
- Character-level attention [25, 26]



[25] Bluche *et al.*, ICDAR 2017

Contribution: Vertical Attention Network (VAN) [19]

Overview



[19] Coquenet *et al.*, TPAMI 2022

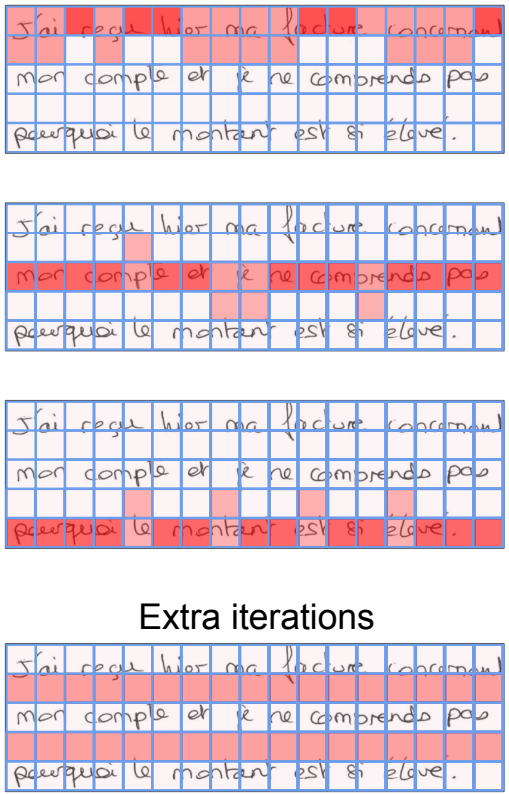
Main contributions

- Line-level vertical hybrid attention
- End-of-paragraph detection module

Vertical Attention Network (VAN)

[24] Bluche *et al.*, NIPS 2016

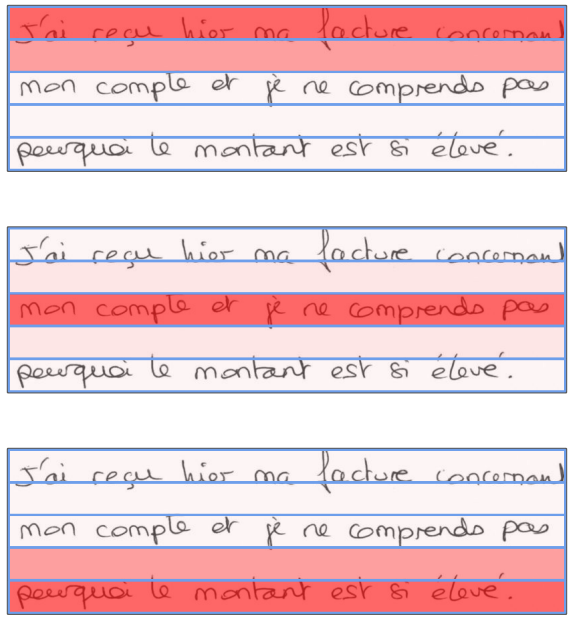
$$\boldsymbol{\alpha}^{t} = g(\boldsymbol{f}, \boldsymbol{\alpha}^{t-1})$$



Fixed number of iterations (*e.g.* T=10)

[19] VAN

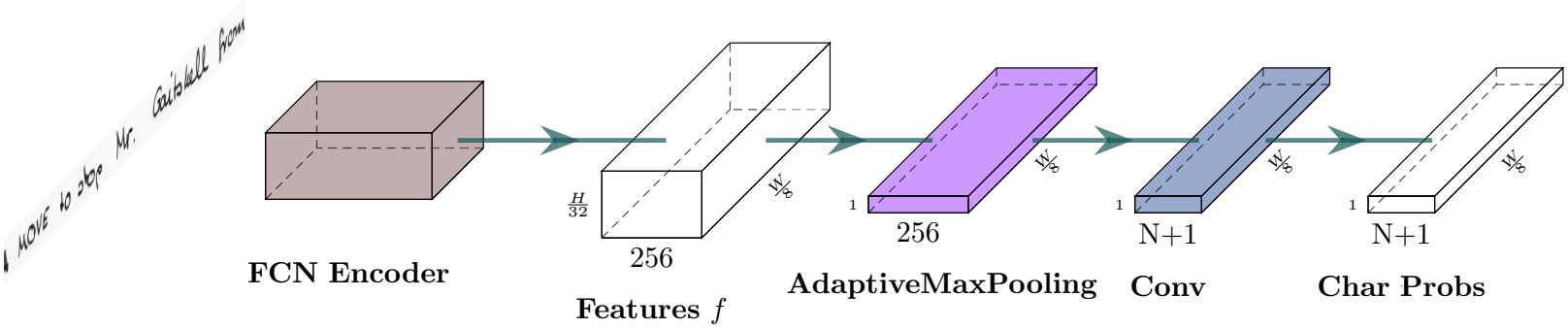
$$\boldsymbol{\alpha}^{t} = g(\boldsymbol{f}, \boldsymbol{\alpha}^{t-1}, \boldsymbol{h}^{W_f \cdot (t-1)})$$

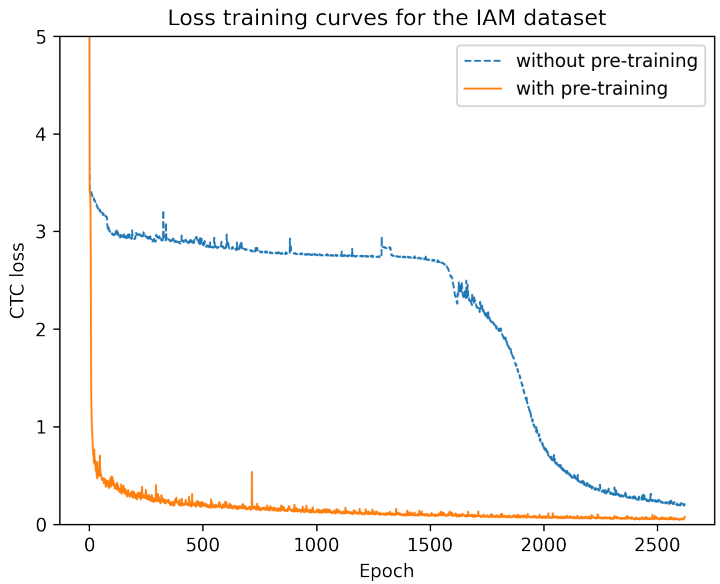


$$\mathcal{L} = \underbrace{\sum_{k=1}^{L} \mathcal{L}_{\text{CTC}}(\boldsymbol{p}^{k}, \boldsymbol{y}^{k})}_{\text{Recognition}} + \lambda \underbrace{\sum_{k=1}^{L+1} \mathcal{L}_{\text{CE}}(\boldsymbol{d}^{k}, \boldsymbol{\delta}^{k})}_{\text{End-of-paragraph detection}}$$

Training strategy

Pre-training on isolated text line images:



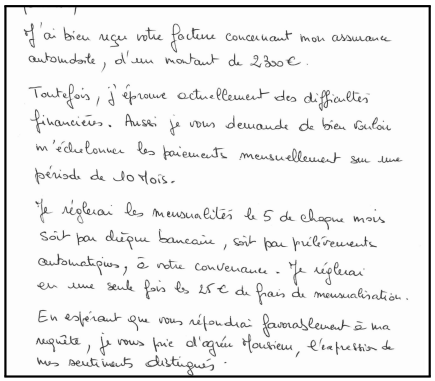


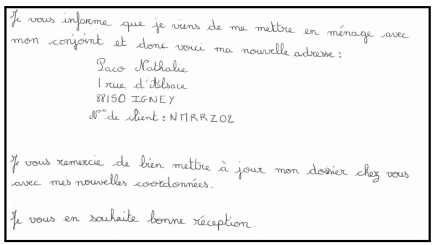
Datasets

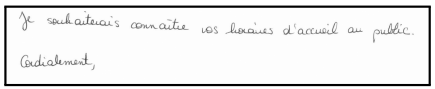
In details

Dataset	Level	Training	Validation	Test	Charset size	Language	# lines
RIMES 2011 [16]	Line	10,532	801	778	100	French	
	Paragraph	1,400	100	100			2-18
IAM [17]	Line	6,482	976	2,915	79	English	
	Paragraph	747	116	336			2-13
READ 2016 [18]	Line	8,349	1,040	1,138	89	Early Modern German	
	Paragraph	1,584	179	197			1-26

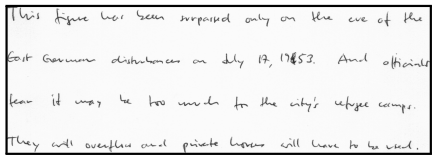
Datasets

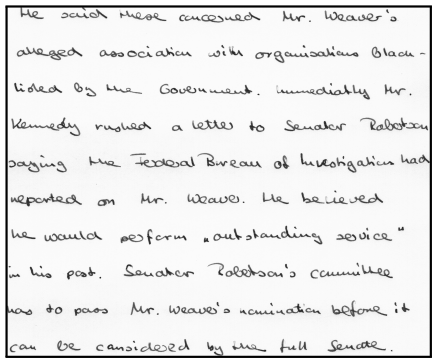


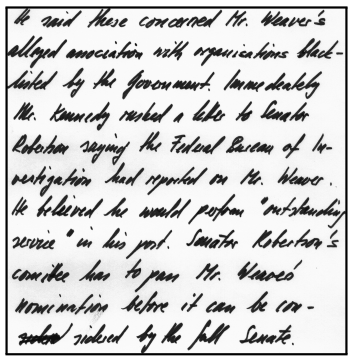




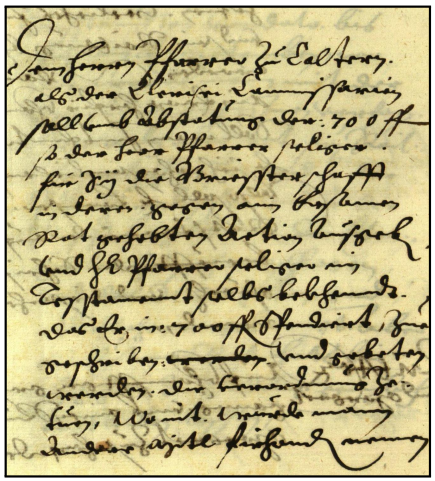
RIMES 2011

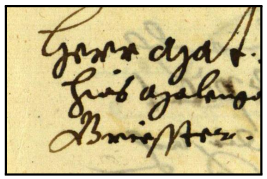


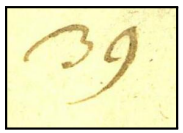




IAM







READ 2016

Paragraph-level recognition results

Paragraph-level state-of-the-art approaches, without language model, external data, nor lexicon constraints.

Architecture	IAM CER (%) test	IAM WER (%) test	RIMES 2011 CER (%) test	RIMES 2011 WER (%) test	READ 2016 CER (%) test	READ 2016 WER (%) test	# Param.
Best line-level approach	4.87[1]		2.3[2]	9.6[2]	4.66[1]		
[25] CNN+MDLSTM[b]	16.2						
[24] CNN+MDLSTM[a]	7.9	24.6	2.9	12.6			
[26] CNN+Transformer[b]	6.7						27.8 M
[23] SPAN (FCN)	5.45	19.83	4.17	15.61	6.20	25.69	19.2 M
[22] OrigamiNet (GFCN)	4.7						16.4 M
[19] VAN (FCN+LSTM)[a]	**4.45**	**14.55**	**1.91**	**6.72**	**3.59**	**13.94**	2.7 M

[1] Results from [14] CNN+BLSTM[b].
[2] Results from [9] CNN+BLSTM.
[a] With line-level attention.
[b] With character-level attention.

VAN demonstration

https://youtu.be/OXi1birmbuw

Conclusion

Bridging the gap between line-level and paragraph-level approaches...

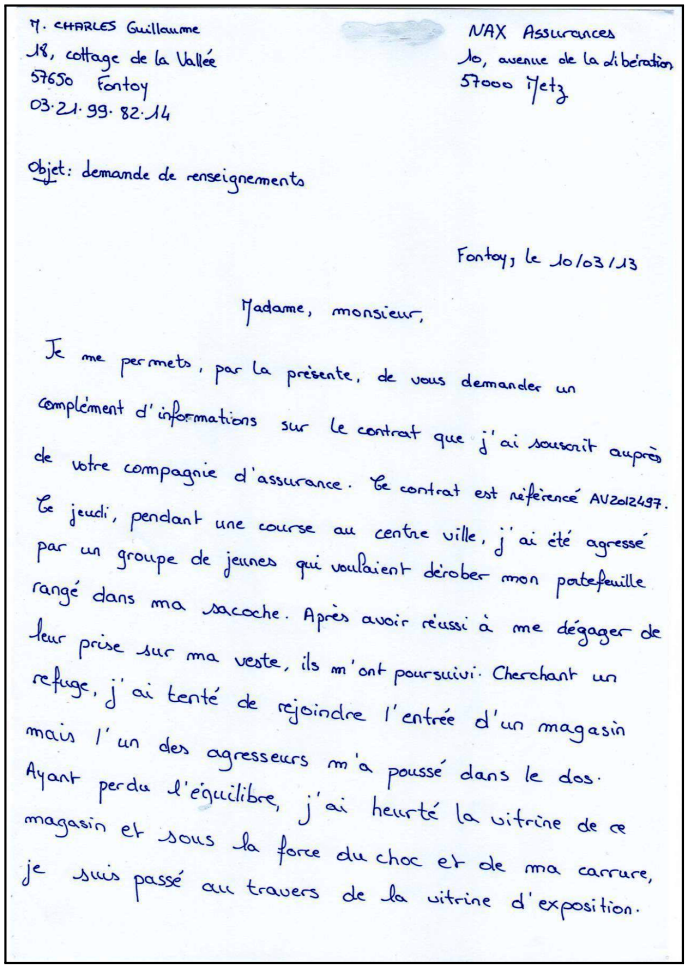
- State-of-the-art results on RIMES 2011, IAM and READ 2016
- Able to deal with slightly inclined lines
- Fast convergence using pre-training
- Few parameters

... but still the same limitations, inherent to the sequential paradigm

- ➤ Rethinking the paradigm

Table of contents

HTR at document level



M. CHARLES Guillaume
18, cottage de la Vallée
57650 Fontoy
03.21.99.82.14

NAX Assurances
10, avenue de la libération
57000 Metz

Objet: demande de renseignements

Fontoy, le 10/03/13

Madame, monsieur,

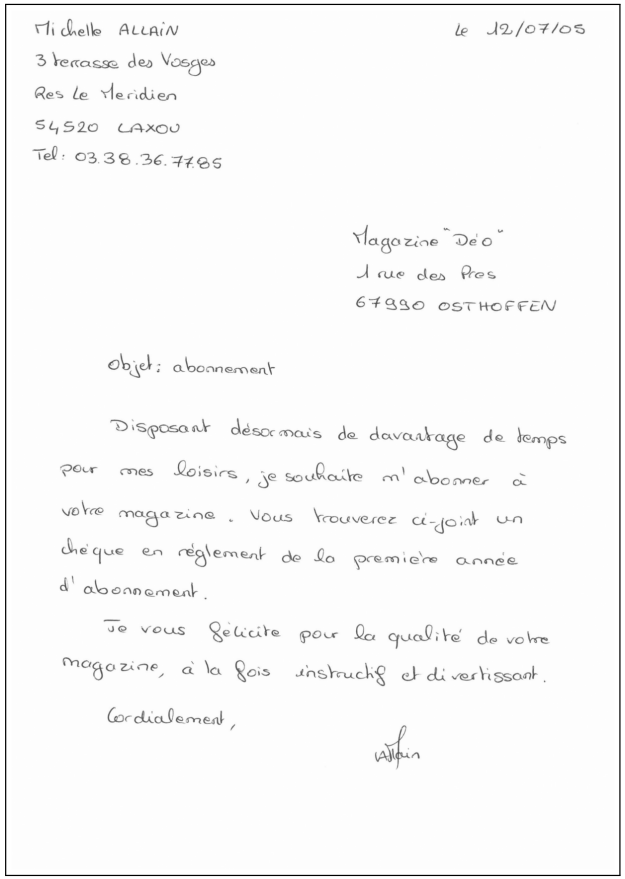
Je me permets, par la présente, de vous demander un complément d'informations sur le contrat que j'ai souscrit auprès de votre compagnie d'assurance. Ce contrat est référencé AV2012497.
Ce jeudi, pendant une course au centre ville, j'ai été agressé par un groupe de jeunes qui voulaient dérober mon portefeuille rangé dans ma sacoche. Après avoir réussi à me dégager de leur prise sur ma veste, ils m'ont poursuivi. Cherchant un refuge, j'ai tenté de rejoindre l'entrée d'un magasin mais l'un des agresseurs m'a poussé dans le dos. Ayant perdu l'équilibre, j'ai heurté la vitrine de ce magasin et sous la force du choc et de ma carrure, je suis passé au travers de la vitrine d'exposition.

Challenges from paragraph to document

- Layout-dependent reading order
- Larger input images and output sequences
  - GPU constraints
  - More complex attention

Handwritten Document Recognition (HDR)

Goal: joint recognition of both text and layout from whole documents

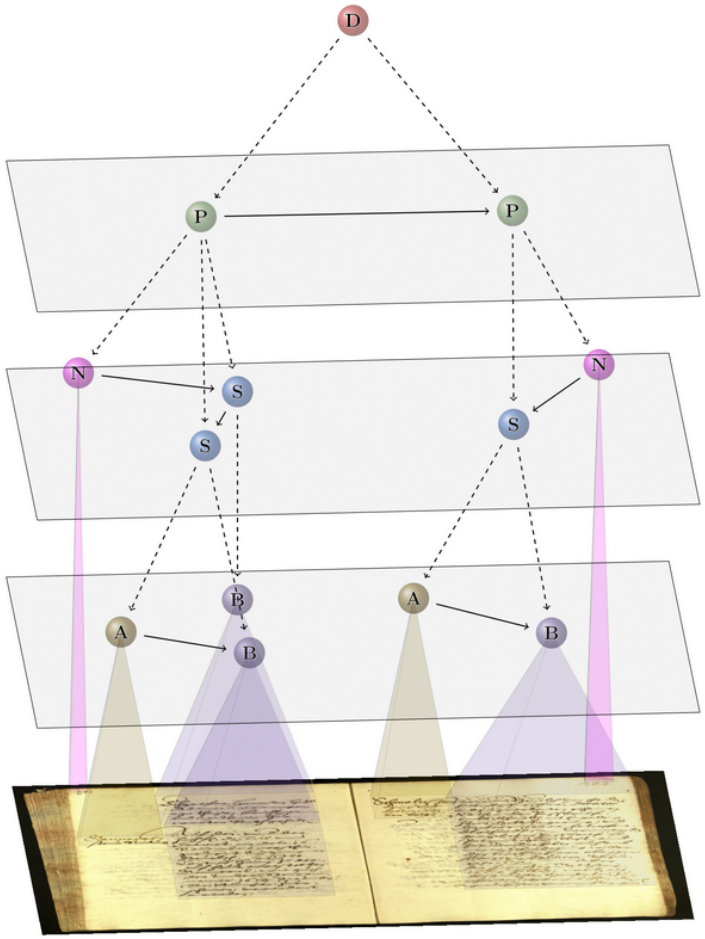


Handwritten Document Recognition

Michelle ALLAIN
3 terasse des Vosges
Res Le Meridien
54520 LAXOU
Tel : 03.38.36.77.85
Le 12/07/05
Magazine "Déo"
1 rue des Pres
67990 OSTHOFFEN
Objet : abonnement
Disposant désormais de davantage de temps pour mes loisirs, je souhaite m'abonner à votre magazine. Vous trouverez ci-joint un chèque en règlement de la première année d'abonnement.
Je vous félicite pour la qualité de votre magazine, à la fois instructif et divertissant.
Cordialement,
Allain

- Sender Coordinates
- Recipient Coordinates
- Place & Date
- Object
- Body
- Signature

How to encode both text and layout ?




```
<document>
  <page>
    <page_number>
      204
    </page_number>
    <section>
      <body>
        Schgrafer, [...] gehalt.
      </body>
    </section>
    <section>
      <annotation>
        Genneral [...] Raitūng
      </annotation>
      <body>
        Aūf den: [...] werden,
      </body>
    </section>
  </page>
  <page>
    <page_number>
      204
    </page_number>
    <section>
      <annotation>
        Schmalz. [...] bet:
      </annotation>
      <body>
        Verer [...] dar¬
      </body>
    </section>
  </page>
</document>
```

- XML paradigm

How to evaluate the performance ?

Evaluate the text recognition

- CER / WER

➤ Normalized edit distance between sequences of characters / words

Prediction: "<A><B>HTR</B>2<B>HDR</B></A>"
Metric computed on: "HTR2HDR"

How to evaluate the performance ?

Evaluate the text recognition

- CER / WER

Evaluate the layout recognition

- LOER (Layout Ordering Error Rate)
- ➤ Normalized edit distance between graphs

Prediction: "<A><B>HTR</B>2<B>HDR</B></A>"
Metric computed on: "<A><B></B><B></B></A>"

How to evaluate the performance ?

Evaluate the text recognition

- CER / WER

Evaluate the layout recognition

- LOER (Layout Ordering Error Rate)

⚠ **Not sufficient:**

Ground truth: "<A><B>HTR</B>2<B>HDR</B></A>"
Prediction: "<A><B></B><B></B></A>HTR2HDR"

LOER = 0% CER = 0%

How to evaluate the performance ?

Evaluate the text recognition

- CER / WER

Evaluate the layout recognition

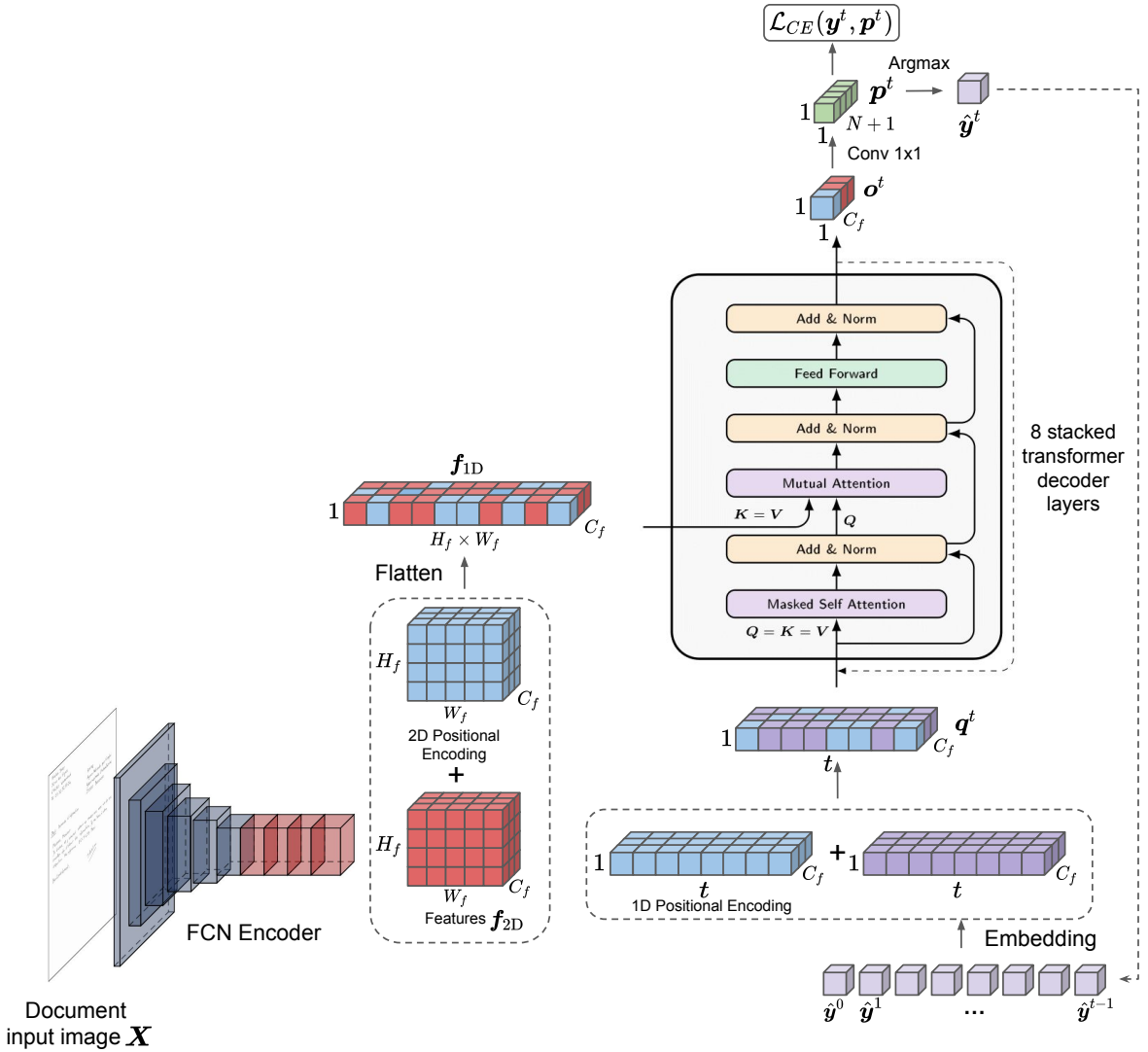
- LOER (Layout Ordering Error Rate)

Evaluate text and layout recognition altogether

- mAP_{CER}
- ➤ Area under the precision / recall curve

Prediction: "<A><B>HTR</B>2<B>HDR</B></A>"
Metric computed on: "HTR2HDR", "HTR", "HDR"

Document Attention Network (DAN) [27]

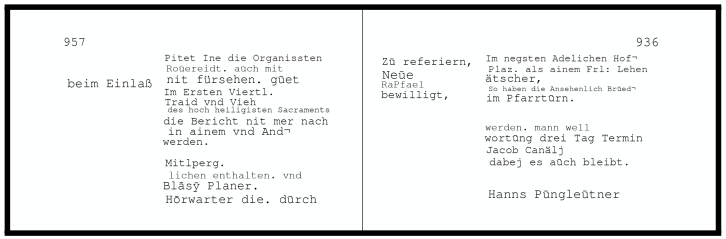


[27] Coquenet *et al.*, Under review 2022

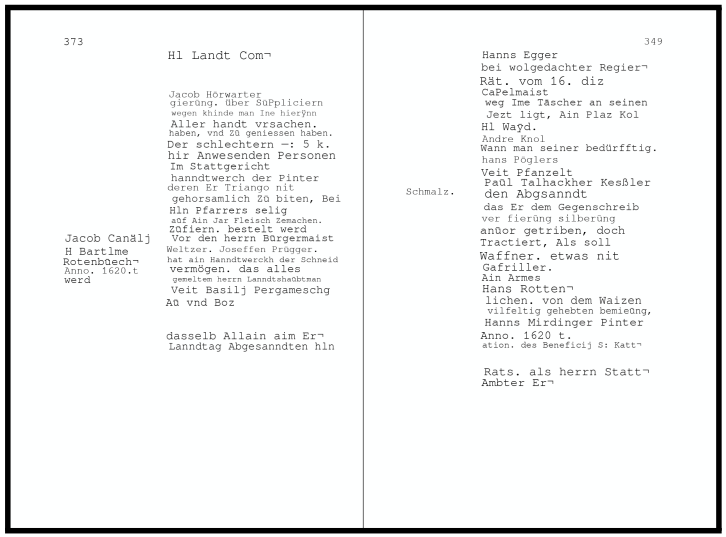
DAN - Training strategy

- Pre-training on synthetic text line images.
- Curriculum learning with synthetic documents:

(a) $l = 3$.

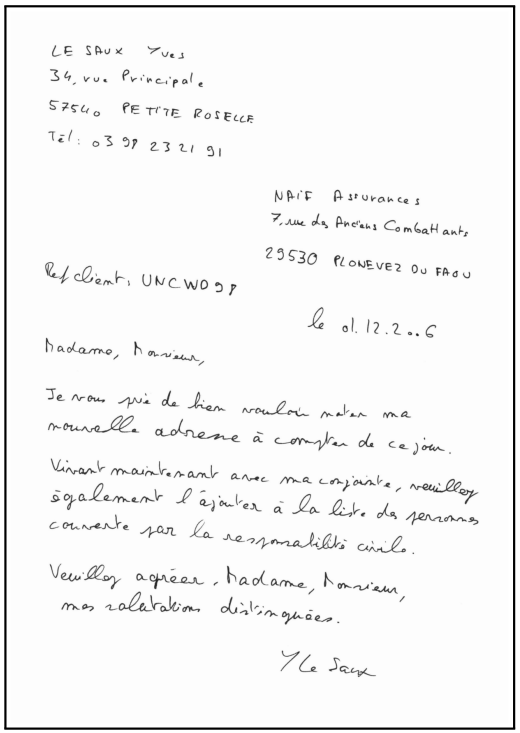


(b) $l = 15$.



(c) $l = l_{\max} = 30$ (end of curriculum stage, no crop).

Datasets


LE SAUX Yves
34, rue Principale
57540 PETITE ROSELLE
Tél: 03 99 23 21 91

NAIF Assurances
7, rue des Anciens Combattants
29530 PLONEVEZ DU FAOU

Ref client: UNCWD 99

le 01.12.2006

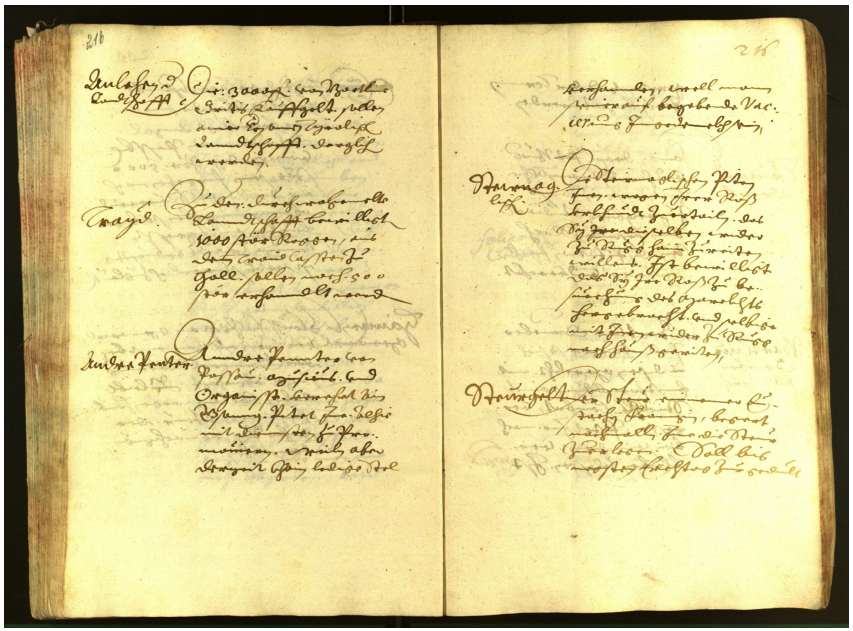
Madame, Monsieur,

Je vous prie de bien vouloir noter ma nouvelle adresse à compter de ce jour.

Vivant maintenant avec ma conjointe, veuillez également l'ajouter à la liste des personnes couvertes par la responsabilité civile.

Veuillez agréer, Madame, Monsieur, mes salutations distinguées.

Y Le Saux



Dataset	Level	Training	Validation	Test	# char tokens	# layout tokens
RIMES 2009 [28]	Page	1,050	100	100	108	14
READ 2016 [18]	Page	350	50	50	89	10
	Double page	169	24	24		

DAN results on the RIMES dataset

⚠ Metrics do not take into account the segmentation step

Dataset	Approach	CER (%) ↓	WER (%) ↓	LOER (%) ↓	mAP_{CER} (%) ↑
RIMES 2011	**Line level**				
	[19] FCN	3.04	8.32	✗	✗
	[9] CNN+BLSTM[a]	**2.3**	9.6	✗	✗
	[27] DAN (FCN+transformer)[c]	2.63	**6.78**	✗	✗
	Paragraph level				
	[23] SPAN (FCN)	4.17	15.61	✗	✗
	[24] CNN+MDLSTM[b]	2.9	12.6	✗	✗
	[19] VAN (FCN+LSTM)[b]	1.91	6.72	✗	✗
	[27] DAN (FCN+transformer)[c]	**1.82**	**5.03**	✗	✗
RIMES 2009	**Paragraph level**				
	[27] DAN (FCN+transformer)[c]	5.46	13.04	✗	✗
	Page level				
	[27] DAN (FCN+transformer)[c]	4.54	11.85	3.82	93.74

[a] This work uses a slightly different split (10,203 for training, 1,130 for validation and 778 for test).
[b] with line-level attention.
[c] with character-level attention.

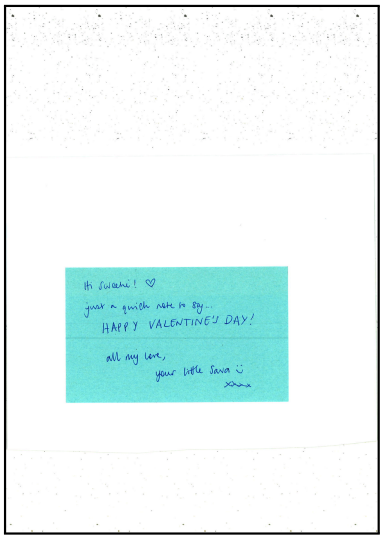
DAN results on the READ 2016 dataset

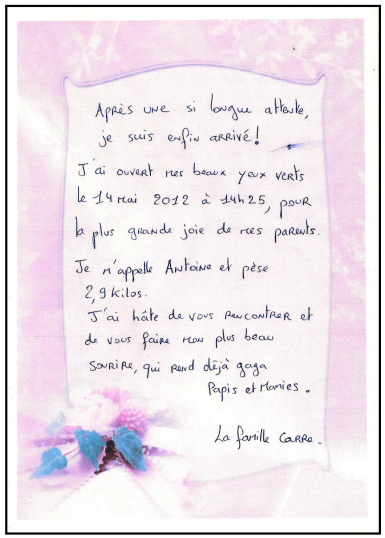
⚠ Metrics do not take into account the segmentation step

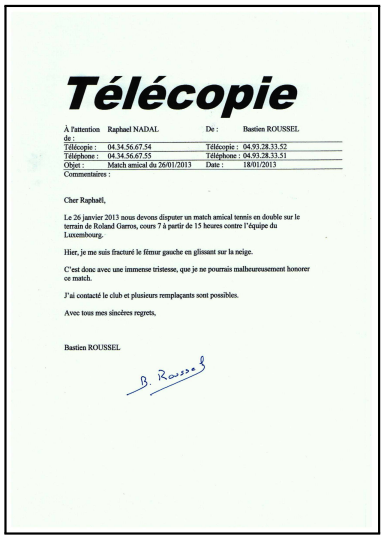
Approach	CER (%) ↓	WER (%) ↓	LOER (%) ↓	mAP_{CER} (%) ↑
Line level				
[14] CNN+BLSTM[a]	4.66	✗	✗	✗
[18] CNN+RNN	5.1	21.1	✗	✗
[19] VAN (FCN+LSTM)[b]	**4.10**	**16.29**	✗	✗
[27] DAN (FCN+transformer)[a]	**4.10**	17.64	✗	✗
Paragraph level				
[23] SPAN (FCN)	6.20	25.69	✗	✗
[19] VAN (FCN+LSTM)[b]	3.59	13.94	✗	✗
[27] DAN (FCN+transformer)[a]	**3.22**	**13.63**	✗	✗
Single-page level				
[27] DAN (FCN+transformer)[a]	3.53	13.33	5.94	92.57
Double-page level				
[27] DAN (FCN+transformer)[a]	3.69	14.20	4.60	93.92

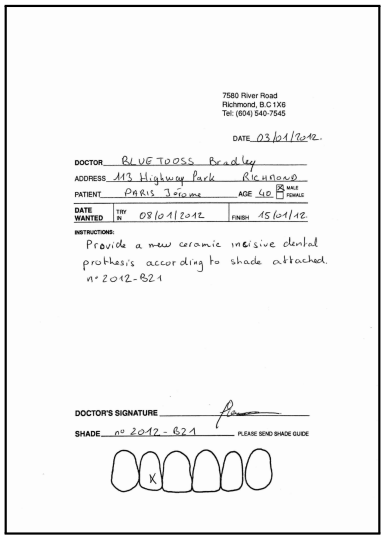
[a] with character-level attention.
[b] with line-level attention.

Experiment on the MAURDOR dataset









Dataset	Training	Validation	Test	# char tokens	# layout tokens	CER (%) ↓ test	WER (%) ↓ test
C3	1,006	148	166	134	✗	8.26	18.94
C4	721	111	114	127	✗	8.02	14.57
C3 & C4	1,727	259	280	141	✗	11.59	27.68

DAN demonstration

https://youtu.be/HrrUsQfW66E

Conclusion

DAN: the first end-to-end model for HDR

- Structured output sequence
- No need for any physical segmentation annotation
- Can follow the slant of the lines (character-level attention)

Line-level / paragraph-level limitations

- ~~Three steps treated independently~~
- ~~A complex pipeline, hard to maintain~~
- ~~Cumulative errors between steps~~
- ~~Additional segmentation annotations~~
- ~~Rule-based reading order~~

Drawback: prediction times grow with the character sequence

Table of contents

General conclusion

Many contributions

Line → Paragraph → Document

Paradigm

From a sequential paradigm for Document Recognition
To a unified paradigm for Document Analysis and Recognition

Attention mechanisms

From text recognition to reading

- Powerful, enable implicit segmentation without annotation
- Require specific training strategies

➤ How to go further ?

Perspectives

Improving the recognition

- Study emerging architectures: Vision Transformer [29, 30, 31]

Dealing with few training data

- Lack of public datasets: self-supervised learning [32, 33, 34]

Reducing the prediction time

- Parallelize the decoding process

[29] Dosovitskiy *et al.*, ICLR 2021
[30] Liu *et al.*, ICCV 2021
[31] Fan *et al.*, ICCV 2021
[32] Caron *et al.*, NIPS 2020
[33] He *et al.*, CVPR 2020
[34] Roh *et al.*, CVPR 2021

Perspectives

Recognizing more

- Handling heterogeneous documents
- Combining HDR with other tasks: Named Entity Recognition, Mathematical Expression Recognition, Table Recognition
- Handling multiple reading orders (schemes, maps)

Towards document understanding

- Document Understanding Transformer for Visual Question Answering [35]

[35] Kim *et al.*, ECCV 2022

Document Understanding

- What is Document Understanding ?
  - Recognition / Analysis ?
  - Key Information Extraction ?
  - Question / Answering ?
  - Inter-document relationship ?
- How to measure the degree of understanding of a document ?
- How to classify the complexity of understanding ?
  (intra/inter-modality: text, table, graph, image, schema, ...)
- What about the connection to the world knowledge ?

Thank you for your attention

References I

[1] Sylvie Brunessaux, Patrick Giroux, Bruno Grilhères, Mathieu Manta, Maylis Bodin, Khalid Choukri, Olivier Galibert, and Juliette Kahn. "The Maurdor Project: Improving Automatic Processing of Digital Documents". In: *International Workshop on Document Analysis Systems (DAS)*. 2014, pp. 349–354.

[2] Guillaume Renton, Yann Soullard, Clément Chatelain, Sébastien Adam, Christopher Kermorvant, and Thierry Paquet. "Fully convolutional network with dilated convolutions for handwritten text line segmentation". In: *International Journal on Document Analysis and Recognition (IJDAR)* 21.3 (2018), pp. 177–186.

[3] Mélodie Boillet, Christopher Kermorvant, and Thierry Paquet. "Robust text line detection in historical documents: learning and Evaluation methods". In: *International Journal on Document Analysis and Recognition (IJDAR)* (2022).

[4] Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks". In: *International Conference on Machine Learning (ICML)*. Vol. 148. 2006, pp. 369–376.

[5] Alex Graves and Jürgen Schmidhuber. "Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks". In: *Advances in Neural Information Processing Systems 21 (NIPS)*. 2008, pp. 545–552.

References II

[6] Vu Pham, Théodore Bluche, Christopher Kermorvant, and Jérôme Louradour. "Dropout Improves Recurrent Neural Networks for Handwriting Recognition". In: *14th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. 2014, pp. 285–290.

[7] Paul Voigtlaender, Patrick Doetsch, and Hermann Ney. "Handwriting Recognition with Large Multidimensional Long Short-Term Memory Recurrent Neural Networks". In: *International Conference on Frontiers in Handwriting Recognition (ICFHR)*. 2016, pp. 228–233.

[8] Curtis Wigington, Seth Stewart, Brian L. Davis, Bill Barrett, Brian L. Price, and Scott Cohen. "Data Augmentation for Recognition of Handwritten Words and Lines Using a CNN-LSTM Network". In: *International Conference on Document Analysis and Recognition (ICDAR)*. 2017, pp. 639–645.

[9] Joan Puigcerver. "Are Multidimensional Recurrent Layers Really Necessary for Handwritten Text Recognition?" In: *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. 2017, pp. 67–72.

[10] Raymond W. Ptucha, Felipe Petroski Such, Suhas Pillai, Frank Brockler, Vatsala Singh, and Paul Hutkowski. "Intelligent character recognition using fully convolutional neural networks". In: *Pattern Recognition* 88 (2019), pp. 604–613.

References III

[11] Denis Coquenet, Yann Soullard, Clément Chatelain, and Thierry Paquet. "Have convolutions already made recurrence obsolete for unconstrained handwritten text recognition ?" In: *Workshop on Machine Learning (WML@ICDAR)*. 2019, pp. 65–70.

[12] Mohamed Yousef, Khaled F. Hussain, and Usama S. Mohammed. "Accurate, data-efficient, unconstrained text recognition with convolutional neural networks". In: *Pattern Recognition* 108 (2020), p. 107482.

[13] Denis Coquenet, Clément Chatelain, and Thierry Paquet. "Recurrence-free unconstrained handwritten text recognition using gated fully convolutional network". In: *17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. 2020, pp. 19–24.

[14] Johannes Michael, Roger Labahn, Tobias Grüning, and Jochen Zöllner. "Evaluating Sequence-to-Sequence Models for Handwritten Text Recognition". In: *International Conference on Document Analysis and Recognition (ICDAR)*. 2019, pp. 1286–1293.

References IV

[15] Christoph Wick, Jochen Zöllner, and Tobias Grüning. "Transformer for Handwritten Text Recognition Using Bidirectional Post-decoding". In: *16th International Conference on Document Analysis and Recognition (ICDAR)*. Vol. 12823. 2021, pp. 112–126.

[16] Emmanuele Grosicki and Haikal El Abed. "ICDAR 2011 - French Handwriting Recognition Competition". In: *International Conference on Document Analysis and Recognition (ICDAR)*. 2011, pp. 1459–1463.

[17] Urs-Viktor Marti and Horst Bunke. "The IAM-database: an English sentence database for offline handwriting recognition". In: *International Journal on Document Analysis and Recognition (IJDAR)* 5.1 (2002), pp. 39–46.

[18] Joan-Andreu Sánchez, Verónica Romero, Alejandro H. Toselli, and Enrique Vidal. "ICFHR2016 Competition on Handwritten Text Recognition on the READ Dataset". In: *15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. 2016, pp. 630–635.

[19] Denis Coquenet, Clément Chatelain, and Thierry Paquet. "End-to-end Handwritten Paragraph Text Recognition Using a Vertical Attention Network". In: *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2022).

References V

[20] Xiao Yang, Ersin Yumer, Paul Asente, Mike Kraley, Daniel Kifer, and C. Lee Giles. "Learning to Extract Semantic Structure from Documents Using Multimodal Fully Convolutional Neural Networks". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 4342–4351.

[21] Yann Soullard, Pierrick Tranouez, Clément Chatelain, Stéphane Nicolas, and Thierry Paquet. "Multi-scale Gated Fully Convolutional DenseNets for semantic labeling of historical newspaper images". In: *Pattern Recognition Letters* 131 (2020), pp. 435–441.

[22] Mohamed Yousef and Tom E. Bishop. "OrigamiNet: Weakly-Supervised, Segmentation-Free, One-Step, Full Page Text Recognition by learning to unfold". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 14698–14707.

[23] Denis Coquenet, Clément Chatelain, and Thierry Paquet. "SPAN: a Simple Predict & Align Network for Handwritten Paragraph Recognition". In: *International Conference on Document Analysis and Recognition (ICDAR)*. Vol. 12823. 2021, pp. 70–84.

References VI

[24] Théodore Bluche. "Joint Line Segmentation and Transcription for End-to-End Handwritten Paragraph Recognition". In: *Advances in Neural Information Processing Systems 29 (NIPS)*. 2016, pp. 838–846.

[25] Théodore Bluche, Jérôme Louradour, and Ronaldo O. Messina. "Scan, Attend and Read: End-to-End Handwritten Paragraph Recognition with MDLSTM Attention". In: *International Conference on Document Analysis and Recognition (ICDAR)*. 2017, pp. 1050–1055.

[26] Sumeet S. Singh and Sergey Karayev. "Full Page Handwriting Recognition via Image to Sequence Extraction". In: *16th International Conference on Document Analysis and Recognition (ICDAR)*. Vol. 12823. 2021, pp. 55–69.

[27] Denis Coquenet, Clément Chatelain, and Thierry Paquet. "DAN: a Segmentation-free Document Attention Network for Handwritten Document Recognition". In: *Under review* (2022). URL: https://arxiv.org/abs/2203.12273.

[28] Emmanuele Grosicki, Matthieu Carré, Jean-Marie Brodin, and Edouard Geoffrois. "Results of the RIMES Evaluation Campaign for Handwritten Mail Processing". In: *10th International Conference on Document Analysis and Recognition (ICDAR)*. 2009, pp. 941–945.

References VII

[29] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *9th International Conference on Learning Representations (ICLR)*. 2021.

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows". In: *International Conference on Computer Vision (ICCV)*. 2021, pp. 9992–10002.

[31] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. "Multiscale Vision Transformers". In: *International Conference on Computer Vision (ICCV)*. 2021, pp. 6804–6815.

[32] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments". In: *Annual Conference on Neural Information Processing Systems (NIPS)*. 2020.

References VIII

[33] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. "Momentum Contrast for Unsupervised Visual Representation Learning". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 9726–9735.

[34] Byungseok Roh, Wuhyun Shin, Ildoo Kim, and Sungwoong Kim. "Spatially Consistent Representation Learning". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 1144–1153.

[35] Geewook Kim, Teakgyu Hong, Moonbin Yim, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. "Donut: Document Understanding Transformer without OCR". In: (2021). URL: https://arxiv.org/abs/2111.15664.

Prediction time

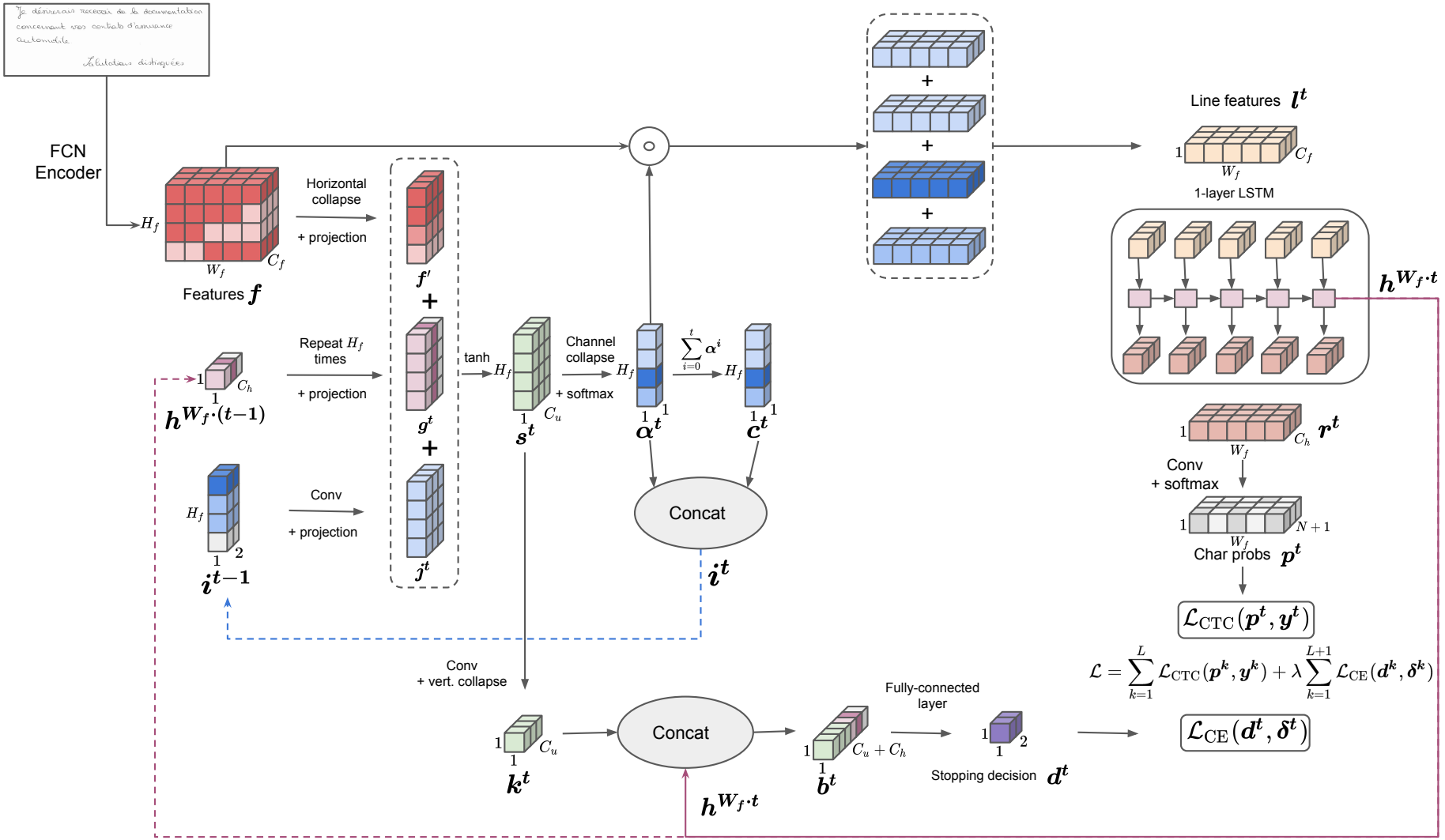
Average prediction time (in seconds) for a test sample, using a single GPU V100 (32Gb).

Dataset	# lines min-max (mean)	# chars min-max (mean)	VAN (FCN+LSTM)[a]	DAN (FCN+transformer)[b]
READ 2016				
Line level	1-1 (1)	1-33 (20)	0.06	0.19
Paragraph level	1-23 (6)	3-547 (117)	0.14	1.43
Single-page level	20-28 (23)	358-583 (468)	✗	4.30
Double-page level	43-51 (46)	831-1,009 (944)	✗	9.70
RIMES 2009				
Paragraph level	1-24 (3)	6-2,043 (104)	✗	1.21
Page level	11-43 (18)	248-2,719 (588)	✗	5.80
RIMES 2011				
Line level	1-1 (1)	5-96 (45)	0.09	0.44
Paragraph level	4-18 (8)	114-944 (359)	0.17	3.15

[a] with line-level attention.

[b] with character-level attention.

Vertical Attention Network (VAN)



VAN - Pre-training strategies

Comparison between cross-dataset pre-training and line-level pre-training for the VAN. Results are given on the test sets.

Source dataset	RIMES CER (%)	IAM CER (%)	READ 2016 CER (%)
Cross-dataset pretraining			
RIMES	✗	4.55	4.08
IAM	1.97	✗	4.14
READ 2016	2.36	5.20	✗
Line-level pretraining			
Target dataset	**1.91**	**4.45**	**3.59**

DAN - Ablation study

	RIMES 2009 (single-page)				READ 2016 (single-page)				READ 2016 (double-page)			
	CER ↓	WER ↓	LOER ↓	mAP_{CER} ↑	CER ↓	WER ↓	LOER ↓	mAP_{CER} ↑	CER ↓	WER ↓	LOER ↓	mAP_{CER} ↑
Baseline	5.72	13.05	**4.18**	**92.86**	**3.65**	14.64	5.51	92.36	4.50	16.75	4.74	**92.37**
(1) No synthetic data	8.26	16.45	8.18	86.34	81.05	94.46	12.04	0.35	80.75	95.65	36.77	0.13
(2) No curriculum for syn. data	7.59	16.48	6.63	88.92	4.28	15.41	5.62	91.66	78.89	92.05	15.42	0.00
(3) No crop in curr. for syn. data	5.84	13.73	4.42	91.94	100.00	100.00	> 100	0.00	100.00	100.00	> 100	0.00
(4) No data augmentation	7.08	15.54	4.78	91.65	4.32	16.67	5.29	91.39	4.92	18.06	5.69	90.92
(5) No curriculum dropout	5.83	14.41	4.36	92.09	3.92	14.85	5.51	**93.13**	**4.23**	**16.12**	**3.68**	92.26
(6) No error in teacher forcing	8.09	15.12	5.91	89.24	7.51	21.87	4.95	83.51	85.78	99.51	42.35	10.73
(7) No layout recognition	**5.30**	**12.46**	✗	✗	4.60	15.59	✗	✗	4.96	16.81	✗	✗
(8) No pre-training	71.42	87.48	18.46	12.72	4.47	16.32	4.72	90.52	5.84	20.47	5.81	88.24
(9) No 1D positional encoding	8.04	16.93	5.73	90.65	3.77	**14.03**	4.95	92.51	4.96	18.28	6.17	88.88
(10) No 2D positional encoding	12.43	20.83	8.42	89.81	5.63	16.25	**4.27**	92.79	65.54	88.43	34.40	25.46

DAN - MAURDOR results in detail

Dataset	Metric	Printed				Hanwdritten				Mix				All			
		FR	EN	Mix	All	FR	EN	Mix	All	FR	EN	Mix	All	FR	EN	Mix	All
C3	# samples	0	0	0	0	42	55	0	97	63	4	2	69	105	59	2	166
	CER (%)	✗	✗	✗	✗	6.13	13.39	✗	8.57	7.86	8.46	10.46	7.98	7.17	12.99	10.46	8.26
	WER (%)	✗	✗	✗	✗	14.83	30.69	✗	20.50	17.10	20.23	25.96	17.50	16.22	29.89	25.96	18.94
C4	# samples	47	9	2	58	0	1	0	1	35	18	2	55	82	28	4	114
	CER (%)	5.39	0.86	10.93	5.05	✗	12.94	✗	12.94	10.67	12.89	12.79	11.26	7.42	9.17	12.01	8.02
	WER (%)	9.94	2.12	12.64	9.05	✗	35.04	✗	35.04	18.36	24.61	23.61	20.45	13.42	17.60	18.98	14.57
C3 & C4	# samples	47	9	2	58	42	56	0	98	98	22	4	124	187	87	6	280
	CER (%)	8.49	0.26	59.83	9.55	6.87	36.01	✗	16.96	9.20	12.59	13.11	9.90	8.51	21.14	27.05	11.59
	WER (%)	13.96	2.95	58.71	14.44	17.84	124.51	✗	56.87	18.42	22.26	24.09	19.23	17.10	68.27	34.52	27.68

DAN - 2D Positional Encoding

$$
\begin{aligned}
\mathrm{PE}_{2\mathrm{D}}(x, y, 2k) &= \sin(w_k \cdot y), \\
\mathrm{PE}_{2\mathrm{D}}(x, y, 2k+1) &= \cos(w_k \cdot y), \\
\mathrm{PE}_{2\mathrm{D}}(x, y, d_{\mathrm{model}}/2 + 2k) &= \sin(w_k \cdot x), \\
\mathrm{PE}_{2\mathrm{D}}(x, y, d_{\mathrm{model}}/2 + 2k + 1) &= \cos(w_k \cdot x), \\
\forall k &\in [0, d_{\mathrm{model}}/4],
\end{aligned}
$$

with

$$
w_k = 1/10000^{2k/d_{\mathrm{model}}}.
$$

We set $d_{\mathrm{model}} = C_f = 256$.

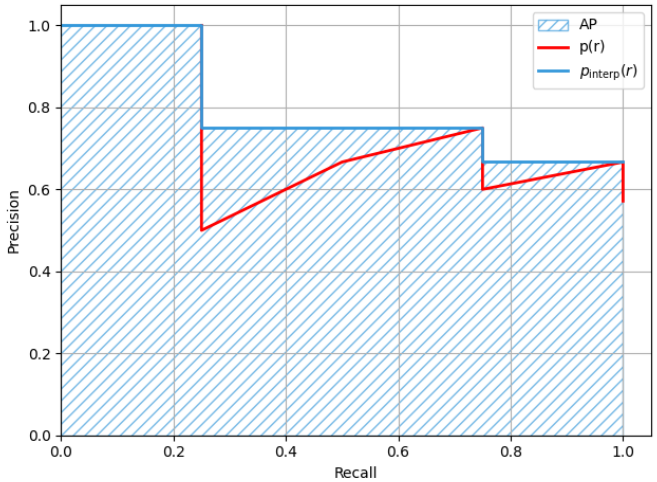
mAP_{CER}

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

$$\text{AP}_{\text{CER}_c} = \sum (r_{n+1} - r_n) \cdot p_{\text{interp}}(r_{n+1}),$$

$$p_{\text{interp}}(r_{n+1}) = \max_{\tilde{r} > r_{n+1}} p(\tilde{r}).$$

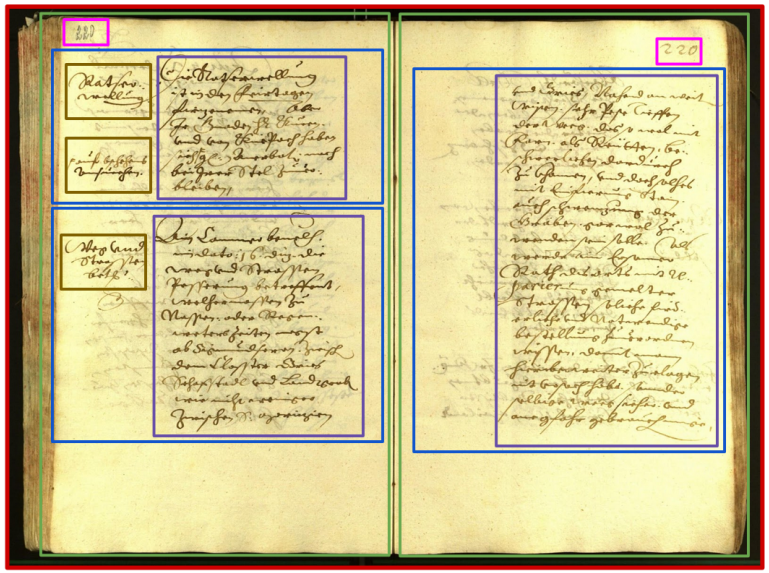


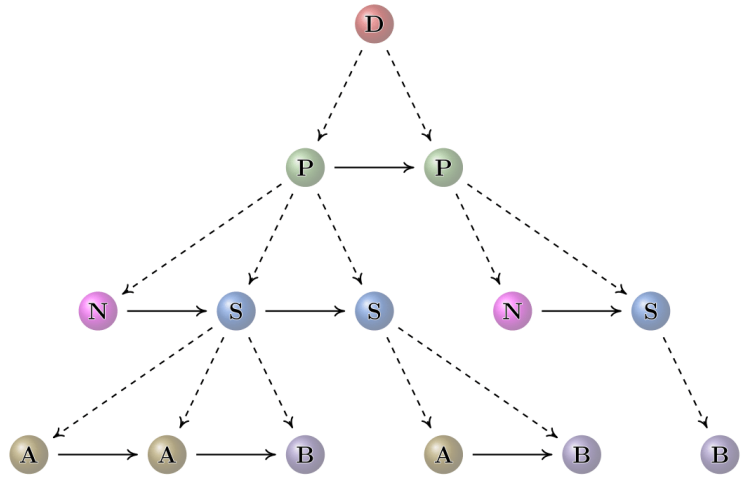
$$\text{AP}^{5:50:5}_{\text{CER}_c} = \frac{1}{10} \sum_{k=1}^{10} \text{AP}^{5k}_{\text{CER}_c}.$$

$$\text{mAP}_{\text{CER}} = \frac{\sum_{c \in S} \text{AP}^{5:50:5}_{\text{CER}_c} \cdot \text{len}_c}{\sum_{c \in S} \text{len}_c}$$

Rank	TP/FP	Precision	Recall	p_{inter}
1	TP	1/1	1/4	1
2	FP	1/2	1/4	1
3	TP	2/3	2/4	3/4
4	TP	3/4	3/4	3/4
5	FP	3/5	3/4	3/4
6	TP	4/6	4/4	4/6
7	FP	4/7	4/4	4/6

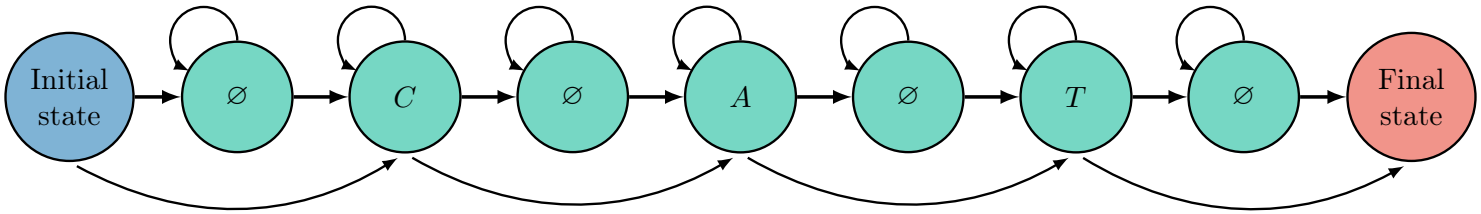
LOER





$$\mathrm{LOER} = \frac{\sum_{i=1}^{K} \mathrm{GED}(\boldsymbol{y}_i^{\mathrm{graph}}, \hat{\boldsymbol{y}}_i^{\mathrm{graph}})}{\sum_{i=1}^{K} n_{\mathrm{e}_i} + n_{\mathrm{n}_i}}.$$

Connectionist Temporal Classification (CTC)



$$\beta(\text{CAAAT}) = \beta(\text{CAT}) = \beta(\text{C}\varnothing\text{AAT}) = \text{CAT}, \text{ but } \beta(\text{CCA}\varnothing\text{AT}) = \text{CAAT}$$

$$\mathcal{L}_{\text{CTC}}(\boldsymbol{p}, \boldsymbol{y}) = -\ln p(\boldsymbol{y}|\boldsymbol{p}).$$

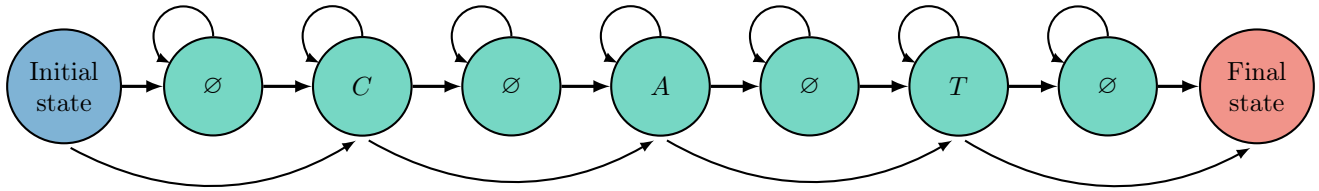
$$p(\boldsymbol{y}|\boldsymbol{p}) = \sum_{\boldsymbol{\pi} \in \mathcal{B}^{-1}(\boldsymbol{y})} p(\boldsymbol{\pi}|\boldsymbol{p}),$$

$$p(\boldsymbol{\pi}|\boldsymbol{p}) = \prod_{t=1}^{L_\text{p}} \boldsymbol{p}^{\boldsymbol{t}}_{\boldsymbol{\pi}^{\boldsymbol{t}}}, \forall \boldsymbol{\pi} \in \mathcal{A}'^{L_p},$$

where $\boldsymbol{p}^{\boldsymbol{t}}_{\boldsymbol{\pi}^{\boldsymbol{t}}}$ is the probability of observing label $\boldsymbol{\pi}^{\boldsymbol{t}}$ at position t in the input sequence $\boldsymbol{p}$.

Connectionist Temporal Classification (CTC)

Ground truth: $\boldsymbol{y} = \text{CAT}$



Prediction $\boldsymbol{p}$:

	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}
C	0.1	**0.9**	**0.8**	0	0.1	0	0.1	0.2	0	0.1
A	0.1	0	0.1	0.2	**0.7**	0.1	0.1	0.2	0.1	0.1
T	0.1	0.05	0.75	0.1	0.1	0.2	0.2	**0.5**	**0.9**	**0.8**
∅	**0.7**	0.05	0.25	**0.7**	0.1	**0.7**	**0.6**	0.1	0	0

Training

$$\mathcal{L}_{\text{CTC}}(\boldsymbol{p}, \boldsymbol{y}) = -\ln p(\boldsymbol{y}|\boldsymbol{p})$$

$$p(\boldsymbol{y}|\boldsymbol{p}) = \sum_{\boldsymbol{\pi} \in \mathcal{B}^{-1}(\boldsymbol{y})} p(\boldsymbol{\pi}|\boldsymbol{p})$$

Evaluation

Best-path decoding: ∅CC∅A∅∅TTT
CTC decoding β:

- rm succ. id. symbols: ∅C∅A∅T
- remove ∅ symbols: CAT

Line-level HTR model (FCN) - Results

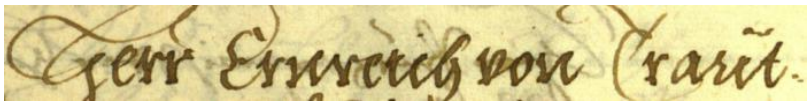
Line-level state-of-the-art approaches, without language model, external data, nor lexicon constraints.

Architecture	IAM		RIMES 2011		READ 2016		# Param.
	CER (%) test	WER (%) test	CER (%) test	WER (%) test	CER (%) test	WER (%) test	
[18] CNN+RNN[a]					5.1	21.1	
[9] CNN+BLSTM	5.8	18.4	**2.3**	9.6			9.3 M
[14] CNN+BLSTM[b]	**4.87**				4.66		
[12] FCN	4.90						> 10 M
[19] Ours (FCN)	5.01	16.49	3.04	8.32	**4.25**	17.14	1.7 M

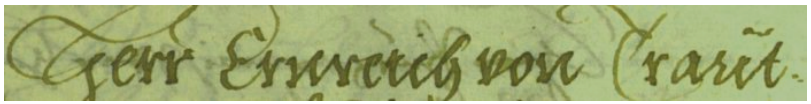
[a] Results from BYU.

[b] With character-level attention.

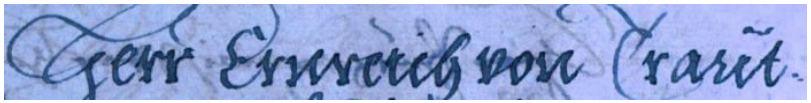
Data augmentation



(a) Original.



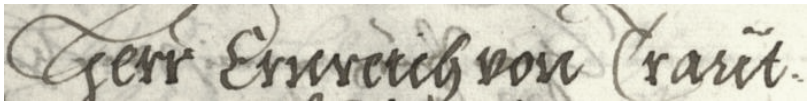
(b) Color jittering.



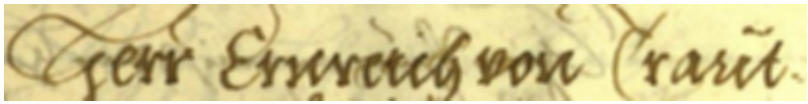
(c) Hue.

(d) Contrast.

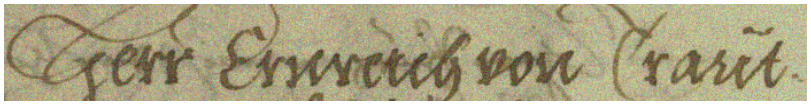
(e) Brightness.



(f) Saturation.

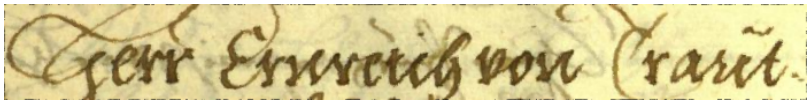


(g) Gaussian Blur.

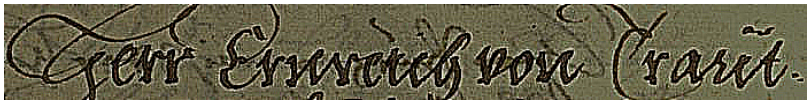


(h) Gaussian Noise.

Data augmentation

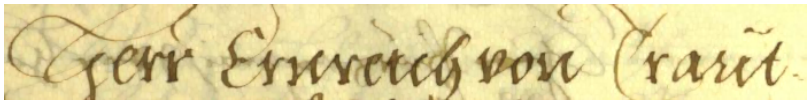


(a) Elastic distortion.

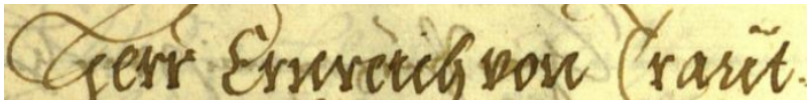


(b) Sharpening.

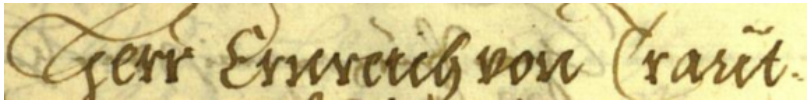
(c) Dilation.



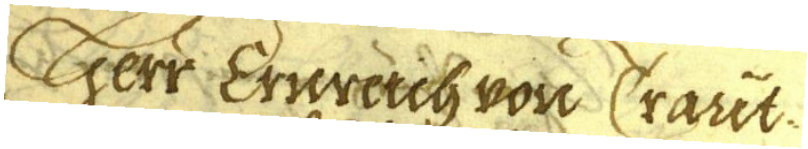
(d) Erosion.



(e) Zoom.



(f) Resolution modification.



(g) Rotation.

(h) Perspective transformation.