

# Text Line Segmentation in Historical Document Images Using an Adaptive U-Net Architecture<sup>+</sup>

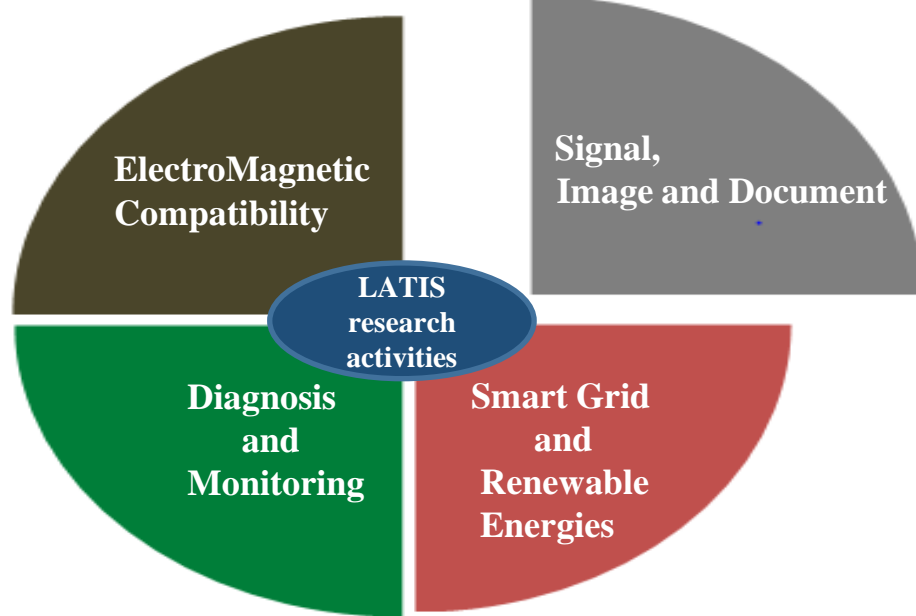
Olfa Mechi<sup>1\*</sup>, Maroua Mehri<sup>1</sup> and Najoua Essoukri Ben Amara<sup>1</sup>

<sup>1</sup>Laboratory of Advanced Technology and Intelligent Systems - ENISo, Tunisia

\* [olfamechi@yahoo.fr](mailto:olfamechi@yahoo.fr)

SIFED'19

## Context and goals



- Collaboration between the Laboratory of Advanced Technology and Intelligent Systems (LATIS, Sousse University)<sup>1</sup> and the Tunisian national archives (ANT)<sup>2</sup>
- Development of reliable tools for :
  - ✓ Text line segmentation
  - ✓ Text recognition
  - ✓ Document indexing

<sup>1</sup> <http://www.latis-eniso.org/>

<sup>2</sup> <http://www.archives.nat.tn/>

**SIFED'19**

## Outline

**1**

**Introduction**

**2**

**Challenges**

**3**

**Proposed framework**

**4**

**Text line segmentation**

**5**

**Experiments**

**6**

**Results**

**7**

**Conclusions and future work**

SIFED'19

## Introduction

# France national archives

Home > Understand

louis

- Louis iane , Place - 6597 documents
- LOUIS , agent - 2086 documents
- LOUIS , Joseph , agent - 1180 documents
- Saint- Louis , Place - 1078 documents
- The Pensec, Louis (1937 ~...), agent - 910 documents
- Port Louis , Place - 824 documents
- Louis , François , agent - 707 documents
- Mexandeau, Louis (1931 ~...), agent - 663 documents
- St. Louis (street) , Place - 655 documents
- LOUIS E - PLACE , Place - 645 documents



For understanding the evolution of our territories, the  
profiles and virtual exhibitions have been made for you.

The [national archives](#) of France hold the [archives](#) of the [7 million archival documents](#) preserved throughout  
France.

You will also find in this section the events celebrated as part of the [national commemorations](#) of the Archives de France.

- [News](#)
- [Explore the archives](#)
- [Genealogy: find my ancestors](#)
- [Archival Publications](#)
- [Thematic guides](#)
- [Cultural and educational offer](#)
- [Commemorate](#)

3,175 documents in this section

# English archives

[Home](#) > [Discovery](#) > [Your search results](#)

[Discovery help](#) [Bookmark](#) [Show quick tour](#)

## Search results for queen

queen



[Start new search](#) [Advanced search](#) [Browse](#)

### Filter results

#### Held by

- [The National Archives \(242,822\)](#)
  - [Available for download only \(104,868\)](#)
- [Other archives \(73,472\)](#)

#### Date

- ☐ [Dates unknown \(18,749\)](#)
- ☐ [1950+ \(21,938\)](#)
- ☐ [1925 - 1949 \(25,513\)](#)
- ☐ [1900 - 1924 \(151,171\)](#)
- ☐ [1800 - 1899 \(74,688\)](#)
- ☐ [1700 - 1799 \(15,086\)](#)
- ☐ [1600 - 1699 \(8,029\)](#)

Records 316,294

Record creators 514

Descriptions of records and information about how to access them

1 - 15 of 316,294 results Sorted by: [Relevance](#) [Simple view](#) [Print](#) [Export results](#)

Sorting is only enabled when there are fewer than 10,000 results. Refine the search to reduce the number of results to enable sorting.

[Harris, Charles \(fl 1686-1705\), diarist of Oxford, diary and account book](#)

diary and account book

Other references: NRA 1097 [Queen's College](#)

Held by: [Oxford University: Queen's College Library](#)

Date: [1686 - 1705](#)

Reference: [None stated](#)

SIFED'19

Introduction

# Tunisian National Archives

Français - English - عربي

Republic of Tunisia  
Presidency of the Government



Home Presentation Virtual visit Funds retained Data base Communication Our partners

## topicality

[Opening hours of the reading room](#)

[Training program of March 13, 2018](#)

[National Archives Day](#)

[Opening hours of the research and consultation room](#)

[2018 training program](#)

[Training Registration Form March 13, 2018](#)

Home » Database

Thakra

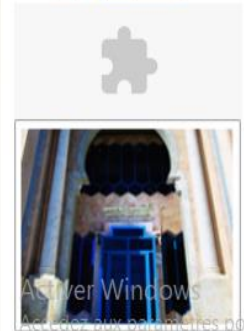
Kiteb

## Thakra

It has about 155,000 records. The unit of description for these documents is at the file level, sometimes at the sub-file level and rarely at the document level (in the case of treaties concluded between Tunisia and foreign countries).

The document description elements are in accordance with ISAD (G) and are:

- ❑ Indications relating to the identification of documents;
- ❑ The indications concerning the content of the documents at the level of the description unit;
- ❑ Extreme dates relating to documents;
- ❑ An indexation that is generally done at three levels:
  - ❑ One or more subject descriptors: there is often a generic descriptor and a specific descriptor;



SIFED'19

Challenges  
(1/4)

## Archival documents of the ANT



Access



Fully automatic  
multilingual text  
transcription  
and document  
indexing system

## Particularities

- Overlapping lines
- Varying-size character
- Problems intrinsic to documents
- Uneven lighting due to folded
- Digitization (page skew, scanning defects)
- Handwritten notes



Categorization of some historical document image particularities [1]

## Particularities

- Arabic script writing styles are varied
- Characters are written in cursive
- Joined right to left
- Capital letters are not comprised
- Characters have 2 to 4 shapes (depending on their location)
- Small space within a word

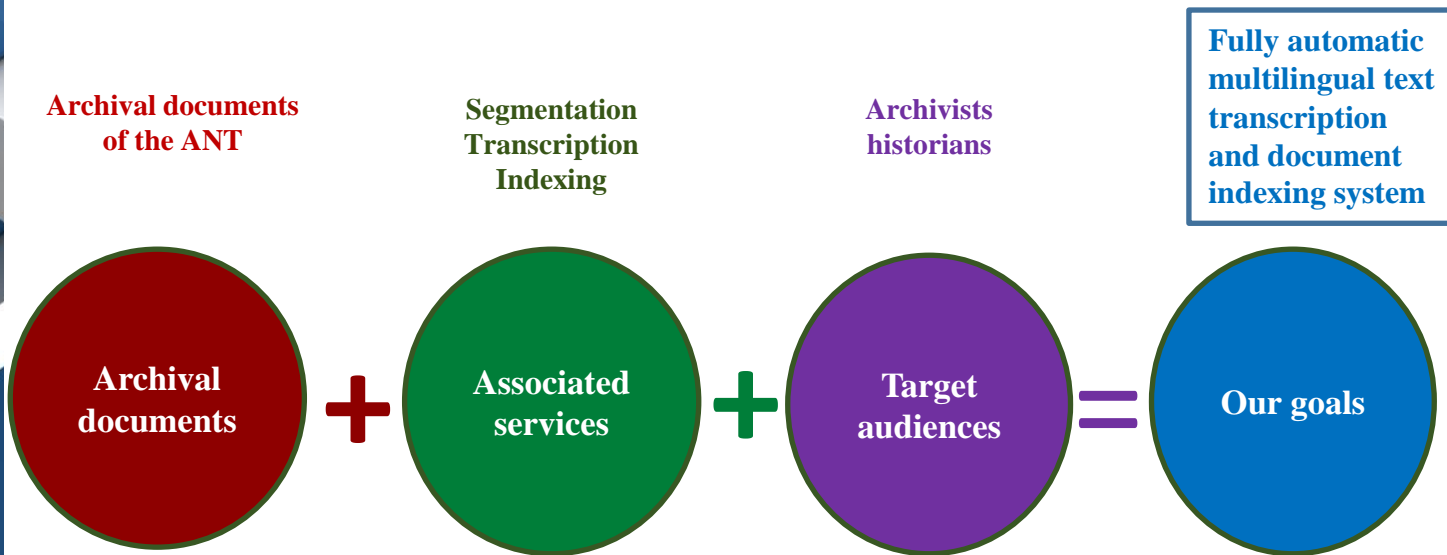
منكرة التخرج لشهادة  
الماستر اختصاص  
مراجعة مناعية

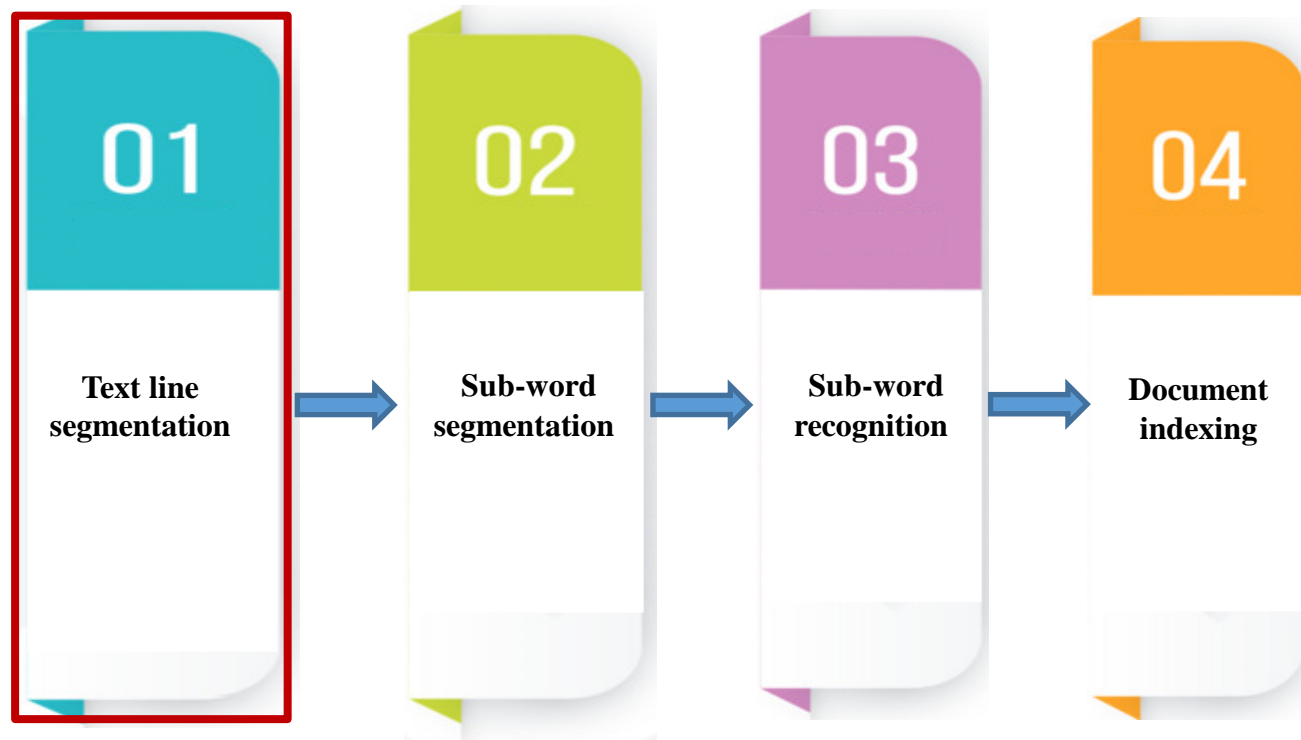
هذه كتابة خطية عربية

Isolate	ع	ع	ع
Initial	ع	ع	ع
Medial	ع	ع	ع
Final	ع	ع	ع

SIFED'19

## Proposed framework (1/2)





Flowchart of the proposed framework

## Text line representations

**Set of pixels:** pixels defined in the textual component [2]

**Enclosing representation** **polygon:** geometric which encloses the total connected text pixels [2]



**Baseline:** virtual line which encloses the most characters whereas descender remains below [3]

**X-height:** area covering the text core, without considering its ascenders and descenders [2]

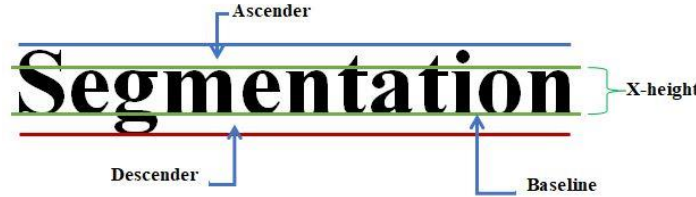
**Illustration of the four representations of a text line used in the literature [2,3]**

[2] Vo *et al.*, "Dense prediction for text line segmentation in handwritten document images", ICIP (2016).

[3] Grüning *et al.*, "READ-BAD: A new dataset and evaluation scheme for baseline detection in archival documents", ICDAR (2017).

SIFED'19

Text line  
segmentation  
(1/6)



**X-height:** area covering the text core, without considering its ascenders and descenders [2]

- Provide additional information useful for the text recognition task (cover the text core)
- Give for each line its associated image  
➡ ease the preparation of the data required for the OCR systems
- Deal with overlapping lines [4, 5]

[2] Vo *et al.*, "Dense prediction for text line segmentation in handwritten document images", ICIP (2016).

[4] Renton *et al.*, "Handwritten text line segmentation using fully convolutional network", ICDAR (2017).

[5] Belabiod and Belaïd, "Line and word segmentation of Arabic handwritten documents using neural networks", HAL (2018).

- Ad-hoc methods: image processing techniques (e.g. connected component, projection, smearing)
  - ✓ Global methods: group components forming the character sequence for each text line, and split components belonging to different text lines [6, 7]
  - ✓ Local methods: determine the local units, and then on group them to define the text lines [8]
  - ✓ Hybrid methods: combine both global and local analysis strategies of text line extraction [9]
- Deep learning-based methods: large number of parameters and layer in neural networks [10, 11]

[6] Shi *et al.*, "A steerable directional local profile technique for extraction of handwritten Arabic text lines", ICDAR (2009).

[7] Alai *et al.*, "A new scheme for unconstrained handwritten text-line segmentation", PR (2011).

[8] Ryu *et al.*, "Language-independent text-line extraction algorithm for handwritten documents", LSP (2014).

[9] Kiumarsi *et al.*, "A hybrid method for text line extraction in handwritten document images", ICFHR (2018).

[10] Barakat *et al.*, "Text line segmentation for challenging handwritten document images using fully convolutional network", ICFHR (2018).

[11] Moysset *et al.*, "Full-page text recognition: learning where to start and when to stop", ICDAR (2017).

## Original U-Net architecture

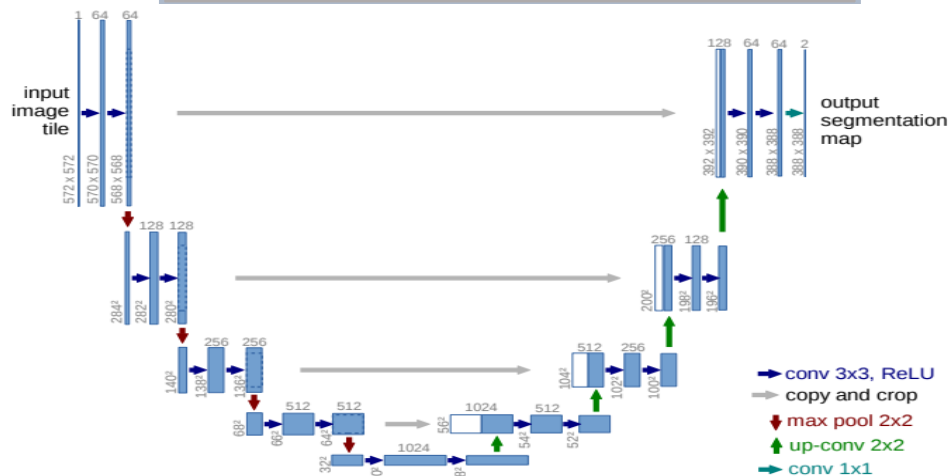


Illustration of the original U-Net architecture [12]

- **U-Net is a variant of FCN**
- **Effectiveness of FCN on semantic segmentation of document images [5, 13]**
- **Computational efficiency**
- **Images of variable sizes can be fed as input of the U-Net architecture**
- **Training phase does not require large volumes of images**

[12] Ronneberger *et al.* "U-Net: convolutional networks for biomedical image segmentation", MICCAI, (2015).

[5] Belabiod and Belaïd, "Line and word segmentation of Arabic handwritten documents using neural networks", HAL (2018).

[13] Grüning *et al.*, "A two-stage method for text line detection in historical documents", CoRR (2018).

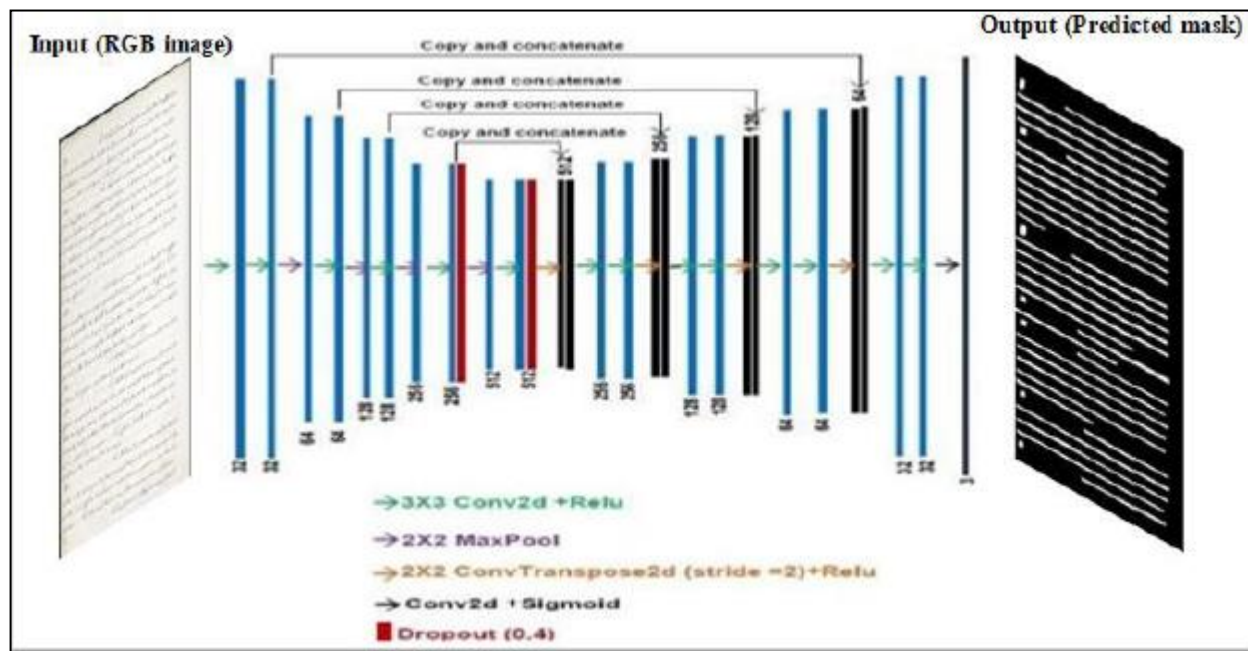


Illustration of the proposed adaptive U-Net architecture

- Reduced number of filters in each convolution layer (**32 filters** at the first block, and then **32, 64, 128, 256 and 512 filters** at each block in the contracting path)
- Use of the deconvolution method "**Conv2DTranspose**" for the decoder phase  
 → in order to keep the same resolution on both the input and output of the network

- The number of parameters in the training phase = **7,760,163** compared with the original U-Net = **31,031,875**
- Reduction of the **memory requirements**, **processing time** and **numerical complexity**
- Compared with the original U-Net architecture, the loss value of the proposed architecture decreases from **0.127** to **0.058** in the training phase
- Elimination of the **over-fitting** issues in the training phase

SIFED'19

Text line  
segmentation  
(6/6)

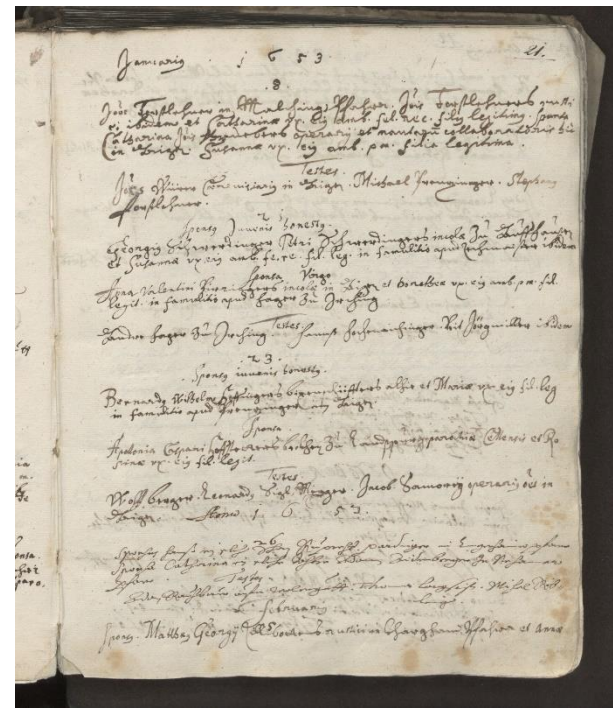
## Experimental corpora

cBAD<sup>5</sup>

DIVA-HisDB<sup>6</sup>

ANT<sup>2</sup>

- Released in the context of the ICDAR competition on baseline detection (cBAD)
- Freely available dataset
- Contains 2,000 handwritten historical document images
- Digitized at 300/400 dpi
- Collected from seven European archives which were drafted between 1470 and 1930



## Experiments (1/4)

<sup>5</sup> <https://scriptnet.iit.demokritos.gr/competitions/5/>

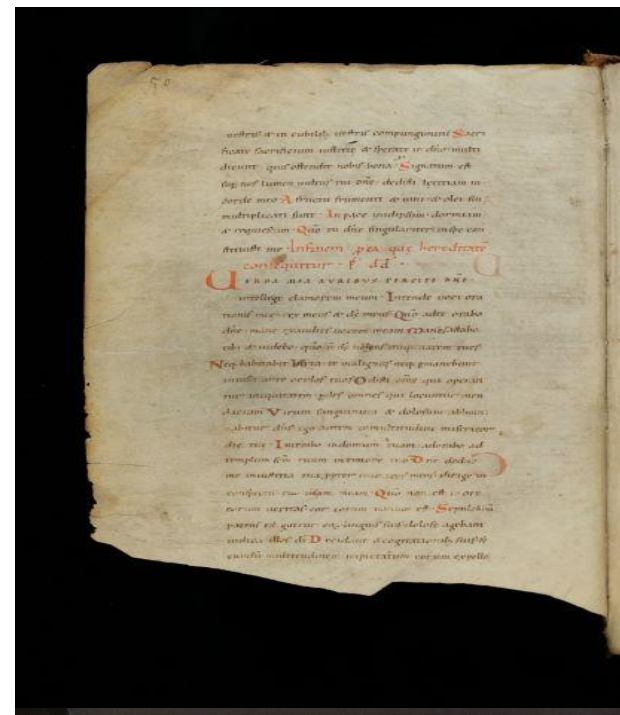
## Experimental corpora

cBAD<sup>5</sup>

DIVA-HisDB<sup>6</sup>

ANT<sup>2</sup>

- Publicly available dataset
- Composed of three medieval manuscripts: CSG18, CSG863 and CB55
- Contains 150 handwritten document images
- Digitized at 600 dpi



Experiments  
(1/4)

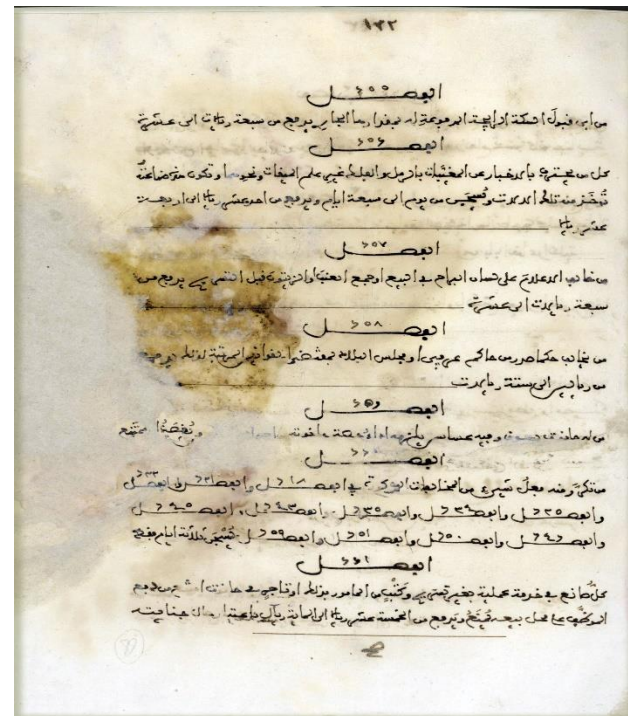
<sup>6</sup> <https://diuf.unifr.ch/main/hisdoc/diva-hisdb>

cBAD<sup>5</sup>

DIVA-HisDB<sup>6</sup>

ANT<sup>2</sup>

- Private corpus collected from the ANT
- Composed of 2 sub-datasets of handwritten document images: *ANT (Arabic)* and *ANT (Latin)*
- 180 and 116 documents written in Arabic and Latin, respectively
- Constitution of the republic of Tunisia
- Written in the 17<sup>th</sup> century
- Digitized at 300 dpi
- Made available as high-resolution color images

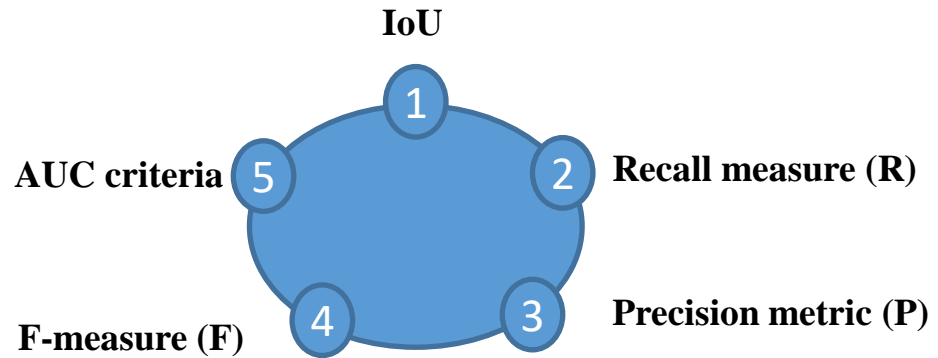


<sup>2</sup> <http://www.archives.nat.tn/>

- The ground truth of the *cBAD* dataset is publicly made available at X-height level
- The X-height annotations of the *READ*, *DIVA-HisDB* and *ANT* datasets are not provided
- We have defined the ground truths of the *DIVA-HisDB* and *ANT* datasets using *Aletheia*<sup>7</sup>
- The ground truths at X-height level of the four datasets (*READ*, *cBAD*, *DIVA-HisDB* and *ANT*) are defined according to the PAGE format [14]

<sup>7</sup> <https://www.primaresearch.org/tools>

[14] Plotschacher and Antonacopoulos, "Page Analysis and Ground-truth Elements) format framework", ICIP (2010).



## Experiments (3/4)

[4] Renton *et al.*, "Handwritten text line segmentation using fully convolutional network", ICDAR (2017).

[15] Diem *et al.*, "cBAD: ICDAR2017 competition on baseline detection", ICDAR (2017).

- Pre-train the network using a private dataset (2499 images) released in the context of the *READ* research project<sup>8</sup>
- Fine-tune the network model on the training sub-dataset of the *cBAD* dataset
- Evaluation on the test sub-datasets: *DIVA-HisDB* (30), *ANT-Arabic* (30), *ANT-Latin* (116) and *cBAD* (539)
- Keras framework
- TITAN X GPU (12 GB allocated memory)
- Learning rate =  $10^{-3}$
- Adam optimizer
- Training images resized to a smaller resolution (608 x 608 pixels)

## Quantitative results

<i>cBAD</i>					<i>DIVA-HisDB (CB55)</i>					<i>ANT (Latin)</i>				
T	P	R	F	IoU	T	P	R	F	IoU	T	P	R	F	IoU
0.1	0.68	0.91	0.77	0.62	0.1	0.49	0.99	0.66	0.47	0.1	0.65	0.97	0.77	0.62
0.2	0.73	0.88	0.78	0.64	0.2	0.51	0.98	0.67	0.49	0.2	0.70	0.96	0.80	0.67
0.3	0.75	0.85	0.79	0.65	0.3	0.52	0.98	0.68	0.50	0.3	0.73	0.94	0.82	0.70
0.4	0.78	0.82	0.79	0.65	0.4	0.53	0.97	0.69	0.52	0.4	0.76	0.93	0.83	0.72
0.5	0.80	0.79	0.78	0.65	0.5	0.55	0.96	0.70	0.53	0.5	0.79	0.91	0.84	0.73
0.6	0.82	0.75	0.77	0.63	0.6	0.56	0.96	0.71	0.54	0.6	0.82	0.88	0.85	0.74
0.7	0.84	0.71	0.75	0.61	0.7	0.58	0.95	0.72	0.56	0.7	0.85	0.85	0.85	0.74
0.8	0.86	0.65	0.73	0.58	0.8	0.61	0.93	0.74	0.58	0.8	0.88	0.80	0.84	0.72
0.9	0.88	0.56	0.67	0.52	0.9	0.66	0.90	0.76	0.61	0.9	0.92	0.71	0.80	0.67
<i>DIVA-HisDB (CSG18)</i>					<i>DIVA-HisDB (CSG863)</i>					<i>ANT (Arabic)</i>				
T	P	R	F	IoU	T	P	R	F	IoU	T	P	R	F	IoU
0.1	0.42	0.94	0.58	0.39	0.1	0.52	0.99	0.68	0.50	0.1	0.50	0.96	0.66	0.48
0.2	0.44	0.92	0.60	0.41	0.2	0.55	0.98	0.71	0.53	0.2	0.54	0.95	0.68	0.51
0.3	0.46	0.90	0.61	0.42	0.3	0.58	0.98	0.73	0.56	0.3	0.56	0.94	0.70	0.53
0.4	0.48	0.88	0.62	0.43	0.4	0.60	0.98	0.75	0.58	0.4	0.59	0.92	0.72	0.55
0.5	0.50	0.86	0.63	0.45	0.5	0.63	0.97	0.76	0.60	0.5	0.61	0.91	0.73	0.57
0.6	0.53	0.84	0.64	0.46	0.6	0.65	0.97	0.78	0.63	0.6	0.63	0.90	0.74	0.58
0.7	0.55	0.81	0.64	0.47	0.7	0.68	0.96	0.79	0.65	0.7	0.65	0.88	0.75	0.59
0.8	0.59	0.76	0.65	0.48	0.8	0.71	0.94	0.81	0.68	0.8	0.68	0.86	0.76	0.61
0.9	0.64	0.69	0.64	0.47	0.9	0.77	0.92	0.83	0.71	0.9	0.72	0.81	0.76	0.62

## Performance evaluation

P, R, F and IoU correspond to the precision, recall, F-measure and intersection over union metrics, respectively.

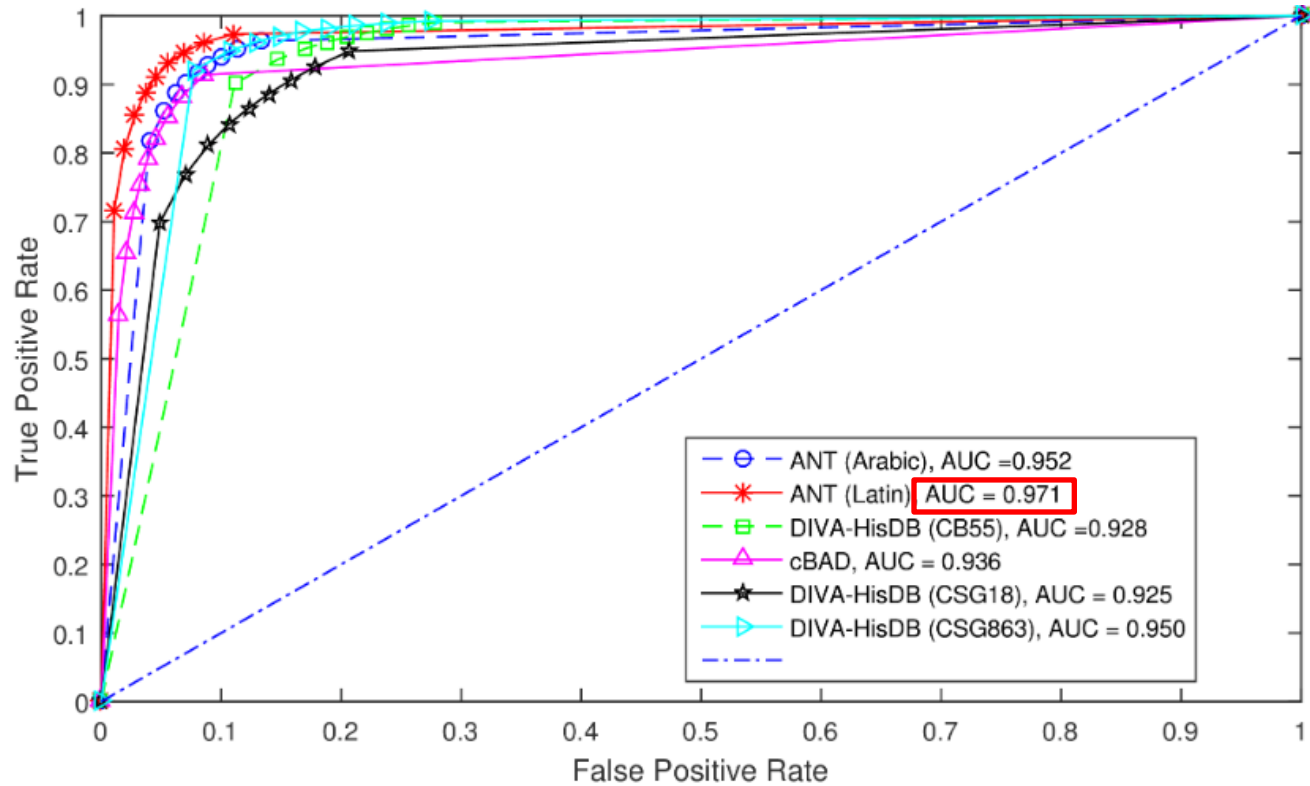
SIFED'19

Results  
(1/6)

SIFED'19

Results  
(2/6)

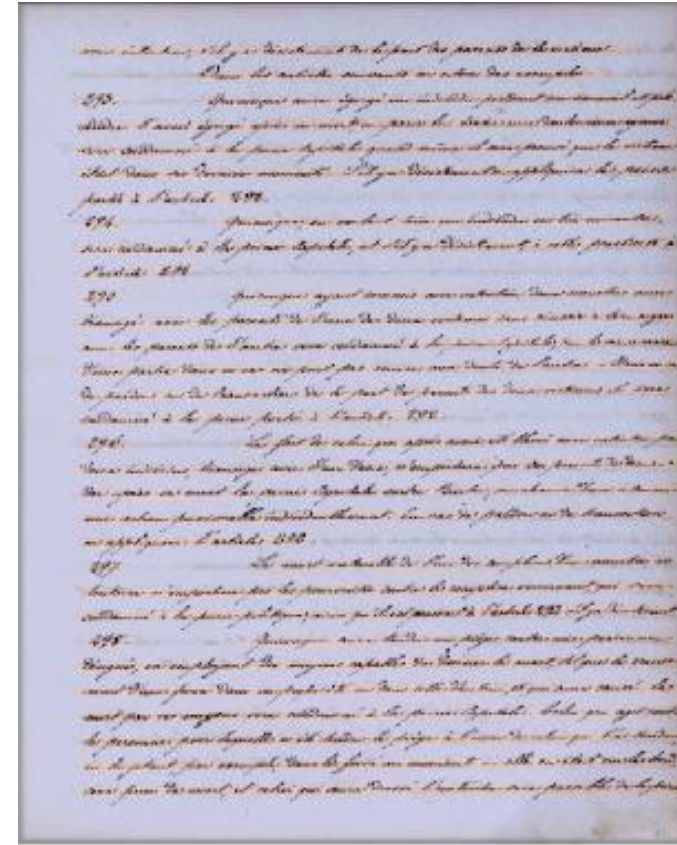
## Quantitative results



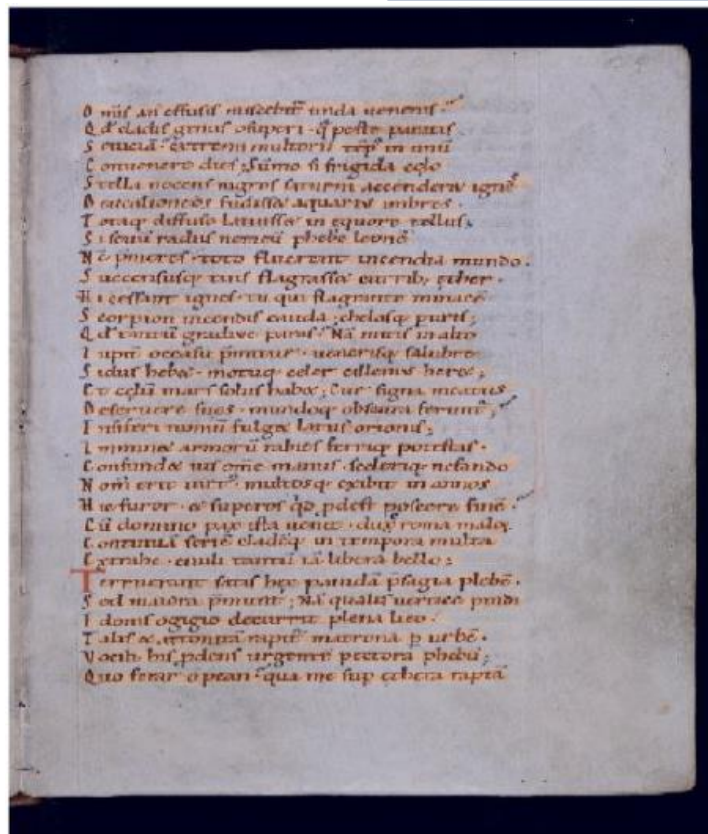
ROC curves of the different assessed datasets

AUC corresponds to the area under the curve

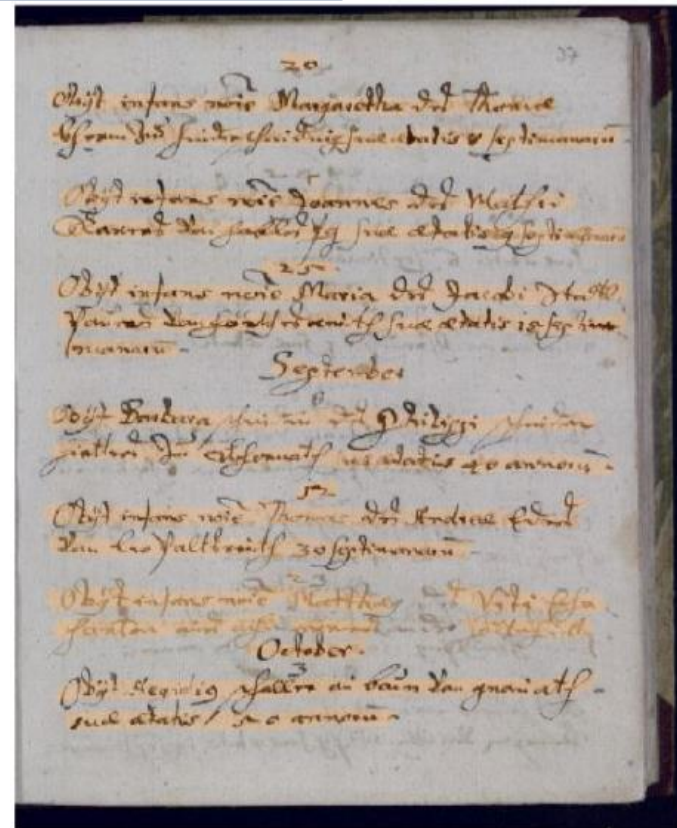
SIFED'19

Results  
(3/6)(a) *ANT (Arabic)*  
(T=0.9, F=76%)(b) *ANT (Latin)*  
(T=0.9, F=85%)

SIFED'19

Results  
(4/6)

(c) *DIVA-HisDB* (CSG863)  
(T=0.9, F=83%)



(d) *cBAD*  
(T=0.3, F=79%)

**TABLE I: Comparison of the performance of the proposed method with Renton *et al.* [4]' method on the *cBAD* dataset**

Method	P	R	F
Proposed method	0.75	0.85	0.79
Renton <i>et al.</i>	0.66	0.86	0.75

**TABLE II: Comparison of the performance of the proposed method with those participating in the ICDAR 2017 Competition on Baseline Detection (*cBAD*) (using baseline as text line representation) [15]**

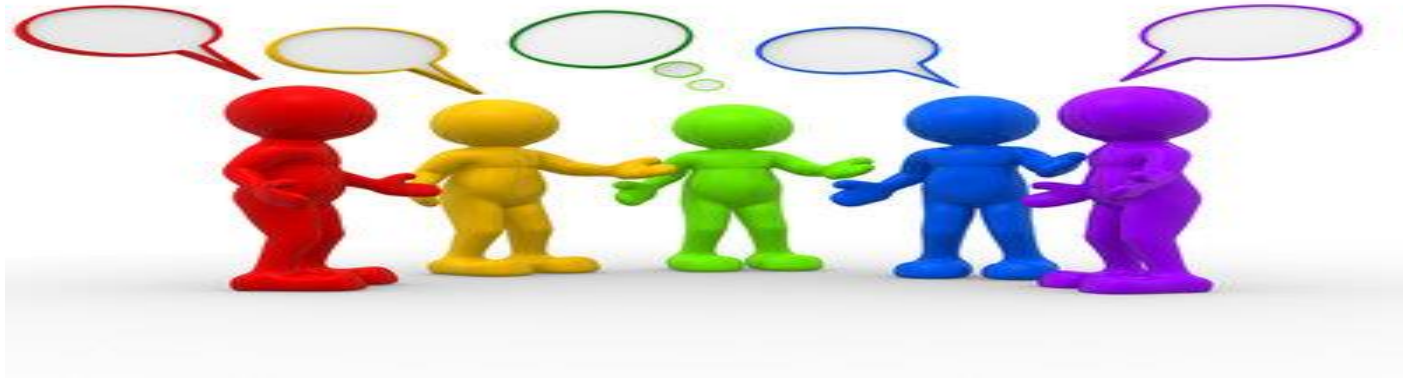
TRACK-A			
Method	P	R	F
DMRZ	0.973	0.970	0.971
UPVLC	0.937	0.855	0.894
BYU	0.878	0.907	0.892
IRISA	0.883	0.877	0.880
LITIS	0.780	0.836	0.807
Proposed method	0.759	0.851	0.791
RU-Net	0.758	0.821	0.774
Dilated FCN	0.745	0.811	0.759

TRACK-B			
Method	P	R	F
DMRZ	0.854	0.863	0.859
BYU	0.773	0.820	0.796
IRISA	0.692	0.772	0.730
Proposed method	0.674	0.805	0.705
UPVLC	0.833	0.606	0.702
RU-Net	0.685	0.738	0.688
Dilated FCN	0.688	0.749	0.669

Results  
(5/6)

[4] Renton *et al.*, "Handwritten text line segmentation using fully convolutional network", ICDAR (2017).

[15] Diem *et al.*, "cBAD: ICDAR2017 competition on baseline detection", ICDAR (2017).



- Compare our method with the participating methods in the ICDAR and ICFHR competitions is not really objective (experimental protocol)
- Integrate pre-processing and post-processing phases based on connected component analysis
  - ➔ correct segmentation errors

1

**Adaptive U-Net architecture  
for segmenting historical  
document into text lines**

2

**Large number of historical  
document images collected  
from different public and private  
benchmarking corpora**

3

**Effectiveness and robustness  
of the proposed architecture  
in both Latin and Arabic  
handwritten document images**

**Method for segmenting the text lines into sub-words**

1

2

**Deep architecture to recognize words in both Arabic and Latin historical document images**

3

**Document indexing system**

**Conclusions  
and future work  
(2/2)**



Thank you  
for your attention



# Questions ?

Contact of first author:

**Olfa Mechi:** [olfamechi@yahoo.fr](mailto:olfamechi@yahoo.fr)

LATIS -- Laboratory of Advanced Technology and Intelligent Systems, ENISo, Tunisia