Centre de Recherche
en Numérique de Sfax
مركز البحث في الرقميات بصفاقس

The Internationl Conference on Intelligent Systems and Pattern Rocognition
16-18 October Hammamet (Tunisia)

Document analysis and recognition : The Deep Learning Era

Yousri Kessentini

Digital Research center of Sfax

yousri.kessentini@crns.rnrt.tn

http://yousri.fr.cr

About me

- Yousri Kessentini
  - Assistant professor at CRNS
  - Head of DeepVision research team
  - More than 15 years experience in the document analysis & recognition
    - PHD in the university of Rouen
    - Post-doc Itesoft-LITIS
    - Researcher in CRNS

Certified as an official instructor & ambassador from NVIDIA Deep Learning Institute

NVIDIA DEEP LEARNING INSTITUTE UNIVERSITY AMBASSADOR

HomePage: http://yousri.fr.cr

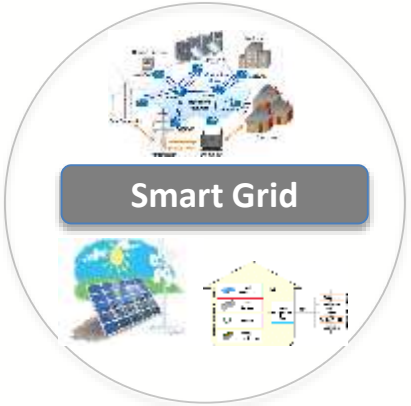
Digital research center of Sfax

- **11 000 m² / 165 researchers**

Research Topics

1. Computer and distributed systems
2. Image and signal processing
3. Document and data analysis
4. Recognition of shapes and objects (3D)
5. Electronics and embedded systems
6. Computer Networks
7. Security of computer systems

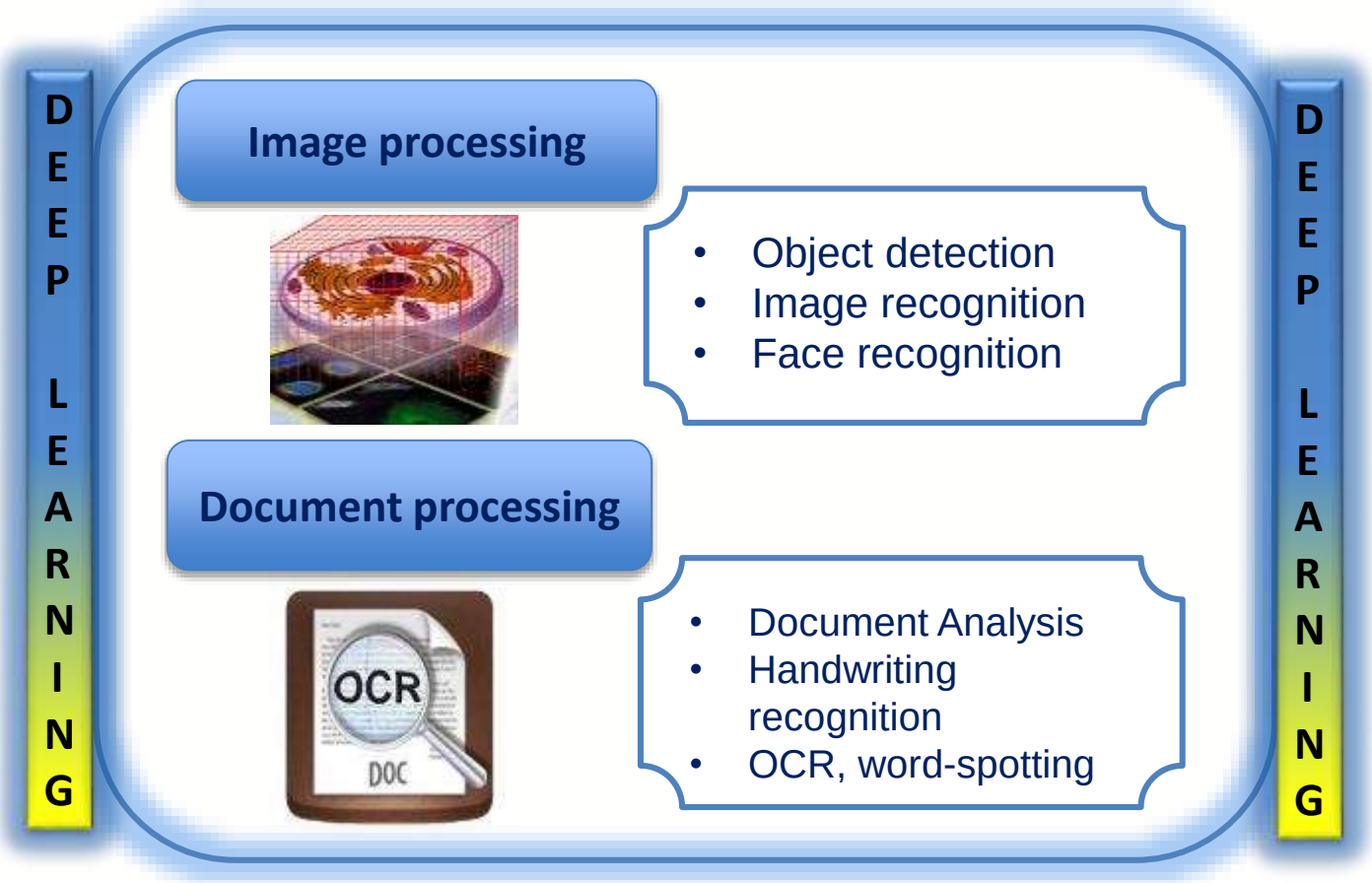
Application Domains



Research activities

DeepVision

Fields : Pattern recognition, Computer vision, Machine learning



Introduction

A huge amount of documents (machine printed, handwritten)

- Preservation
- Storage
- Access to contents

Administrative Documents

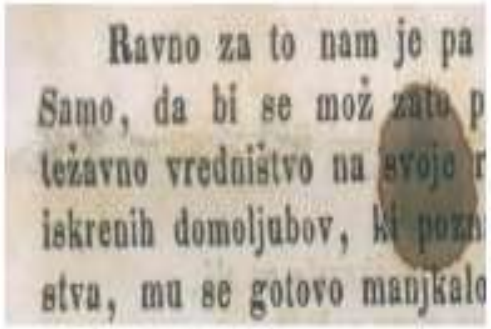
Historical Documents

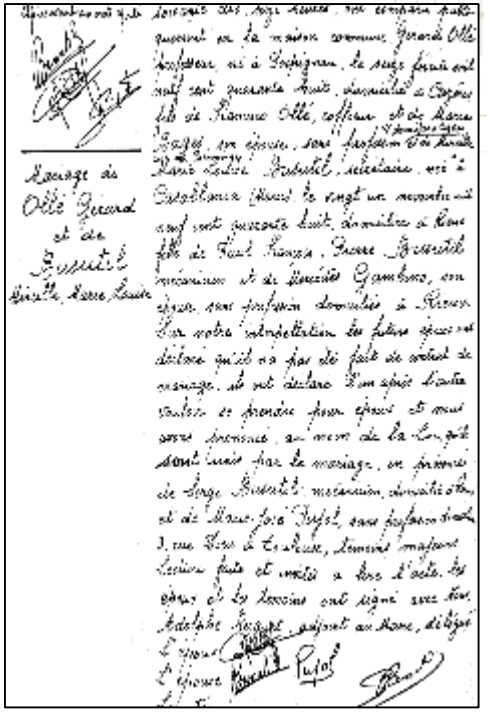
➔ High Cost of manual processing

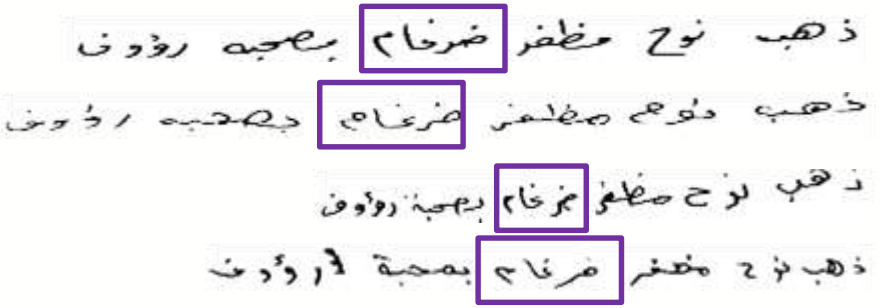
PayStream Advisors research reveals the average **cost** of **manually processing** an invoice can be as high as $20, versus $4 for automated invoice

Introduction

- Automatic document processing
  - Classification
  - Document Enhancement
  - Writer/script Identification
  - Layout analysis
  - Recognition
  - Spotting …
- Difficulties
  - Unstructured data
  - Deformations, noise
  - Different handwriting styles
  - Segmentation problem

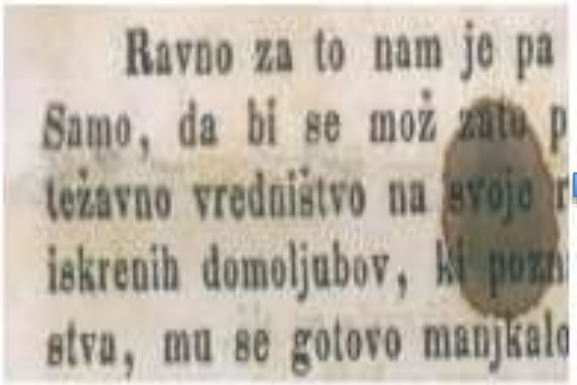




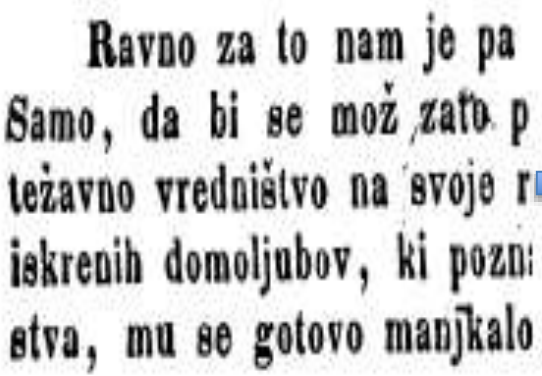


Document Recognition process

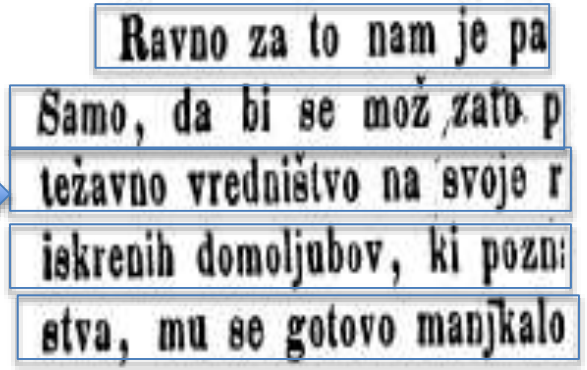
Original document



Document enhancement



Line segmentation

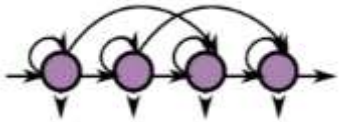


Script identification

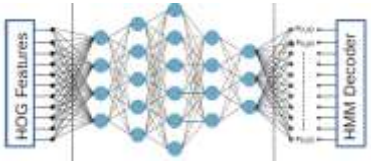
Document recognition

Preprocessing
Feature Extraction
Optical Model
Langage Model

HMM



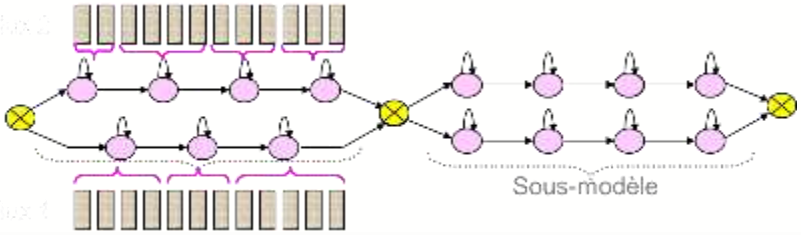
NN-HMM



Deep Learning

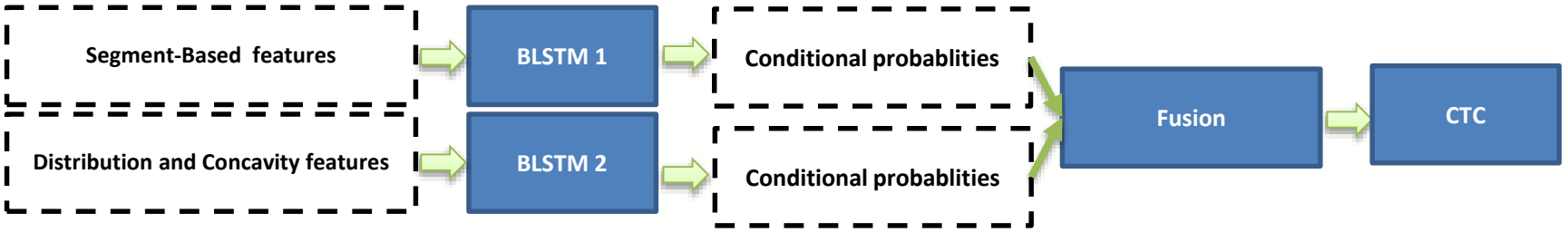
Handwriting recognition before deep Learning

- GMM-HMM with carefully chosen features



Y. Kessentini, T. Paquet, A. Benhamadou. Off-Line Handwritten Word Recognition Using Multi-Stream Hidden Markov Models. Pattern recognition Letters (PRL), Vol 30, Issue 1, pp. 60-70, January 2010.

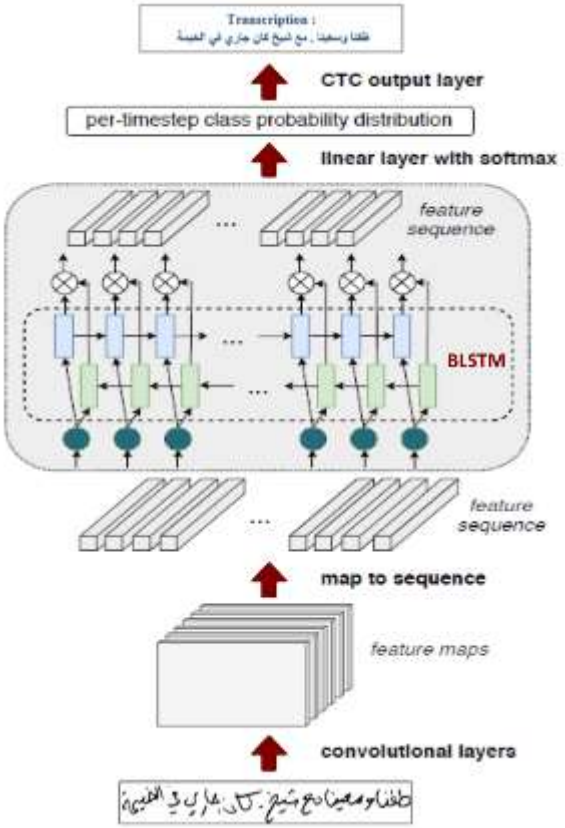
- Combination of BLSTMs at different levels



S. Khamekhem Jemni, **Y. Kessentini**, S. Kanoun, Offline Arabic Handwriting Recognition Using BLSTMs Combination. 13th international workshop on document analysis systems (DAS), pp 31-36, Vienna, Austria, 2018.

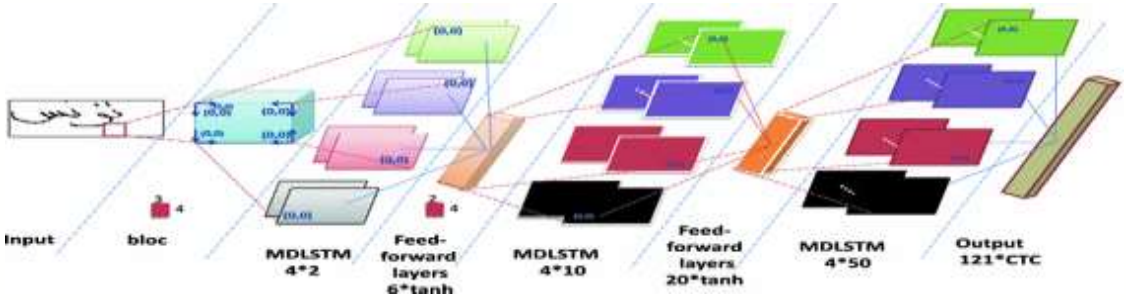
Deep Learning arrived

- Applied directely to the pixel of the raw text line image
  - **BLSTM → MDLSTM → CRNN → Seq2Seq …**

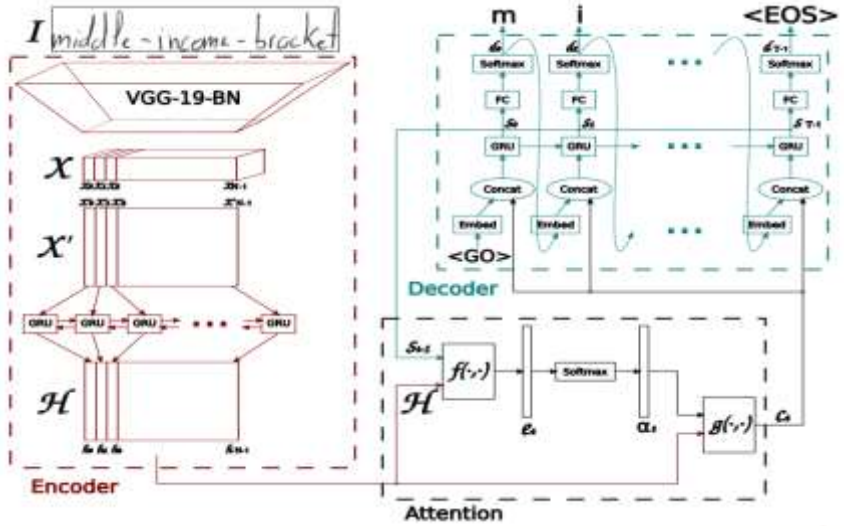


CRNN : CNN+BLSTM

MDLSTM



Seq2Seq model

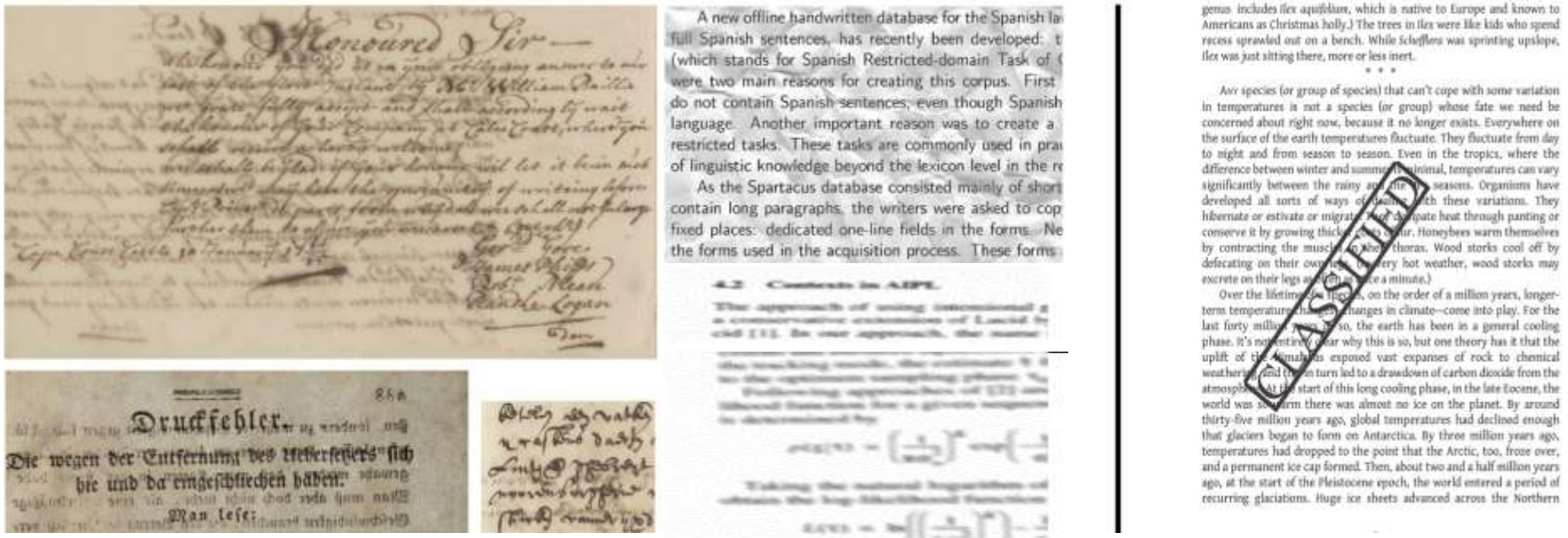


HIGH PERFORMANCE WITH DL
BUT…

How to recognize degraded documents ?

Document Enhancement

- Use Generative Adversarial Networks (GAN) to restore severely degraded document images.



M. A. Souibgui, **Y. Kessentini**, DE-GAN: A Conditional Generative Adversarial Network for Document Enhancement, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), September 2020

Components of GANs

The discriminator tries to identify real data from fakes created by the generator

The generator turns noise into an imitation of the data to try to trick the discriminator

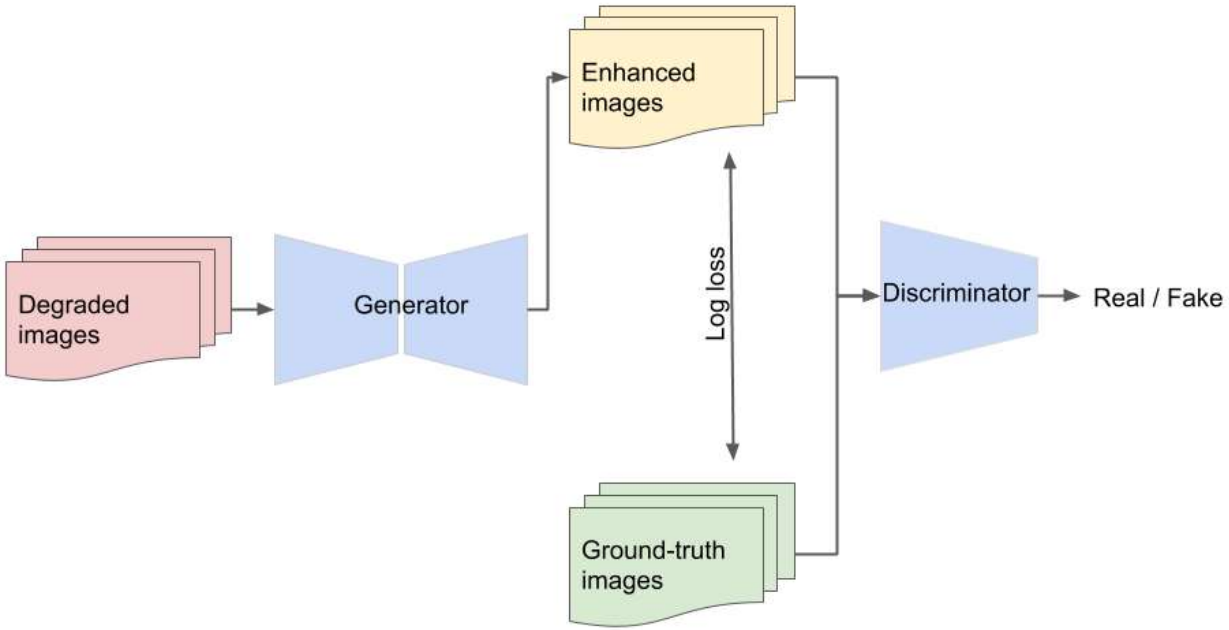
Train GAN jointly via minimax game:

- The discriminator tries to maximize its classification accuracy
- The generator tries to minimize the discriminator's classification accuracy

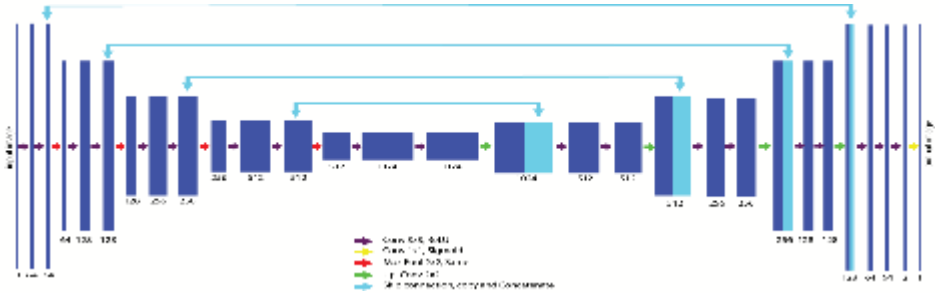
$$\min_G \max_D V(D,G) \quad V(D,G) = \mathbb{E}_{x\sim p(x)}[\log D(x)] + \mathbb{E}_{z\sim q(z)}[\log(1 - D(G(z)))]$$

Maximized by D Minimized by G

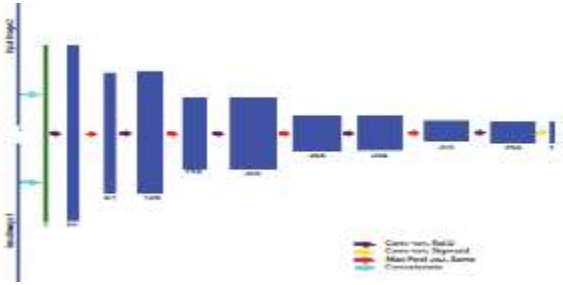
DE-GAN : Document Enhancement GAN



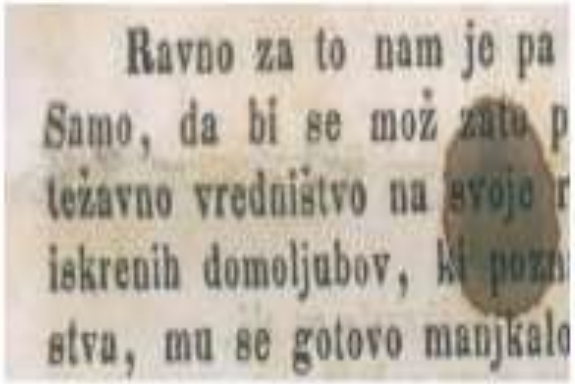
Generator



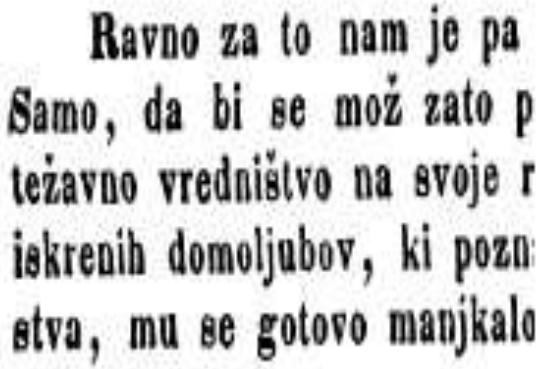
Discriminator



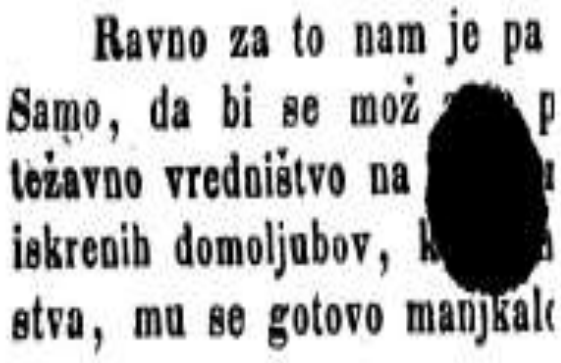
DE-GAN : Document Enhancement



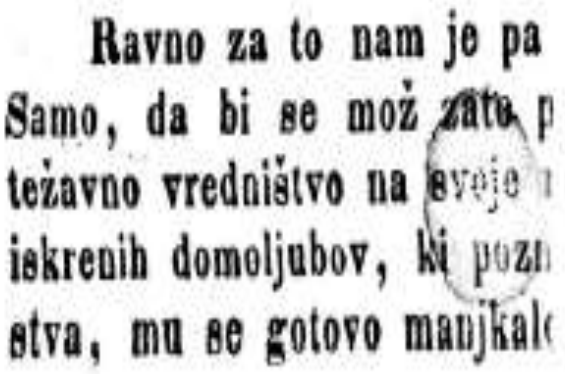
Original



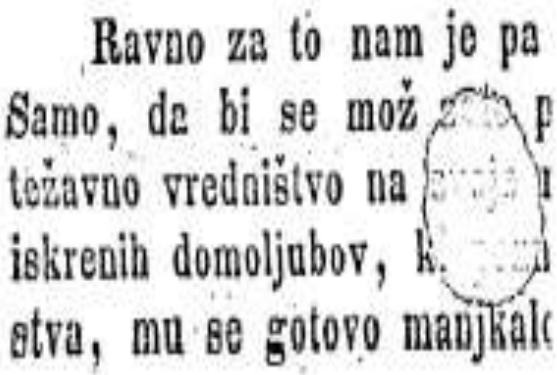
Ground truth



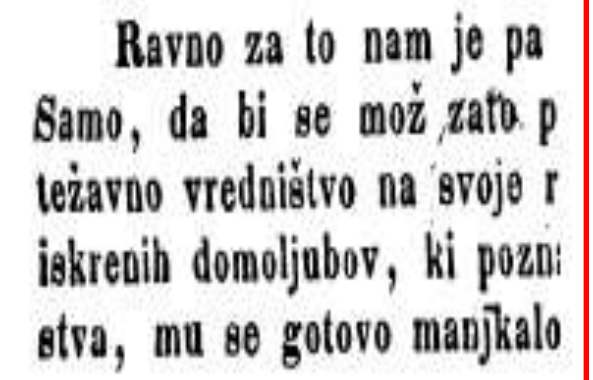
Otsu [16]



Niblack [50]

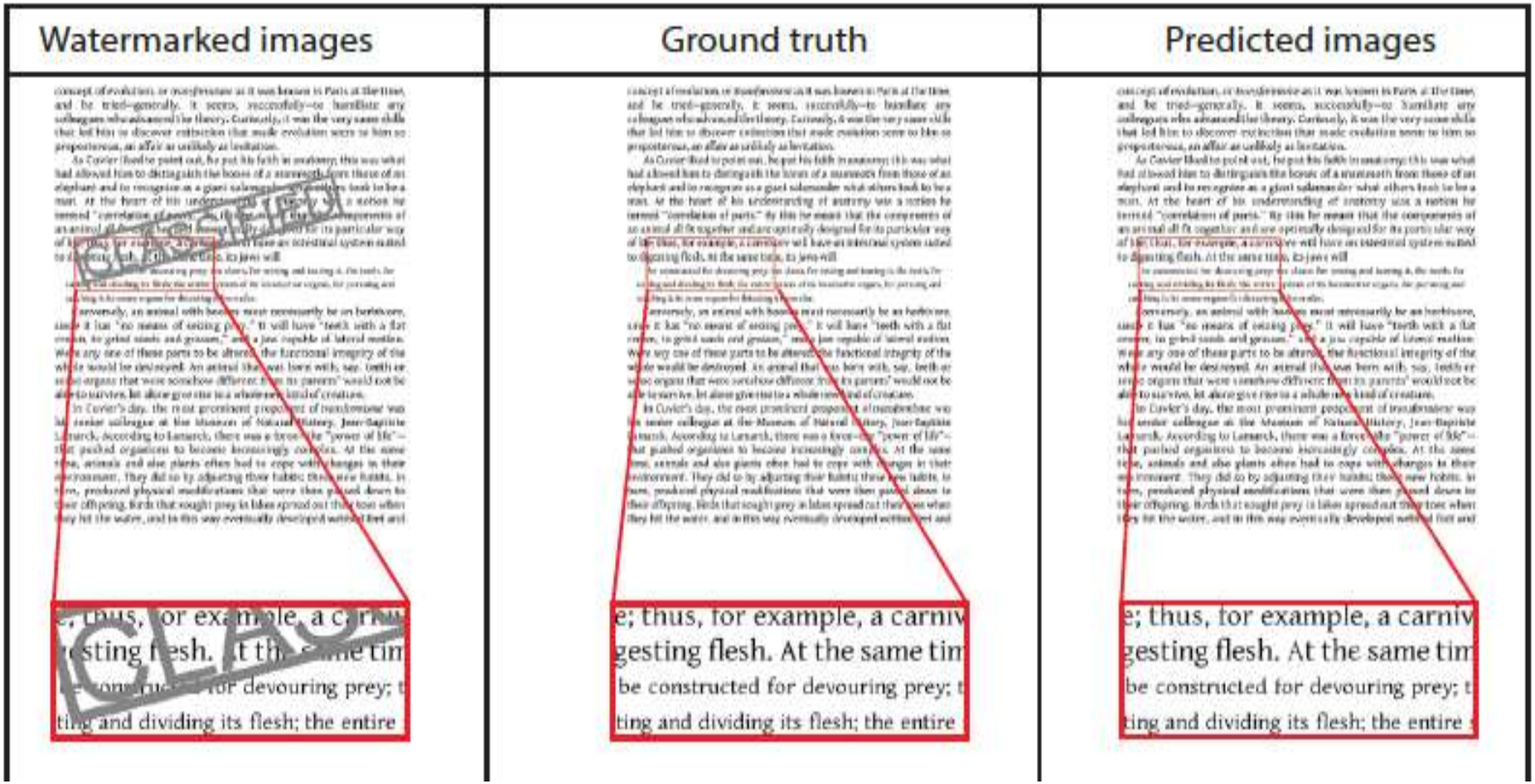


Sauvola et al. [51]

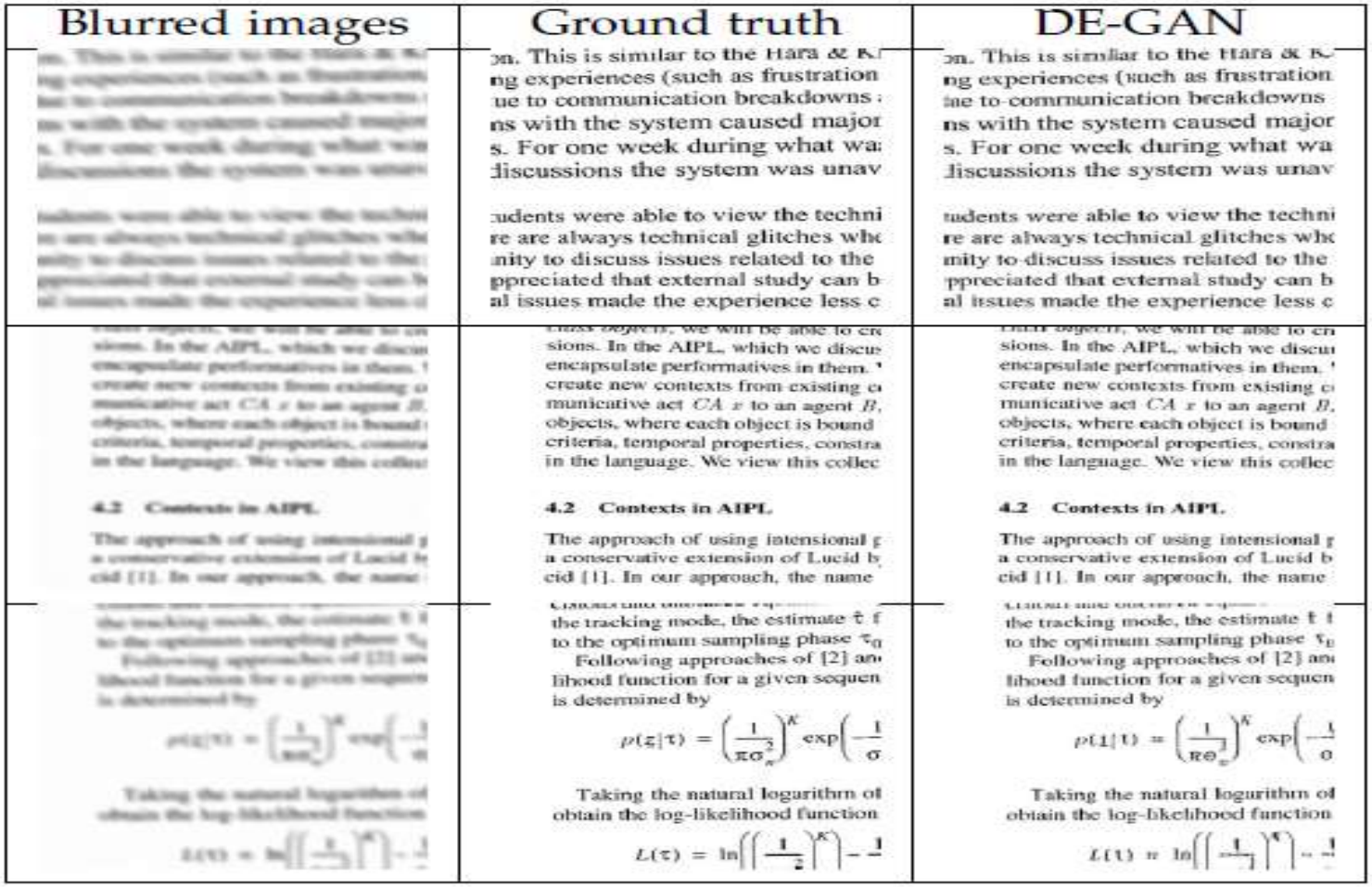


DE-GAN

DE-GAN : Watermark removal



DE-GAN : Debluring



Source code: https://github.com/dali92002/DE-GAN

HIGH PERFORMANCE WITH DL

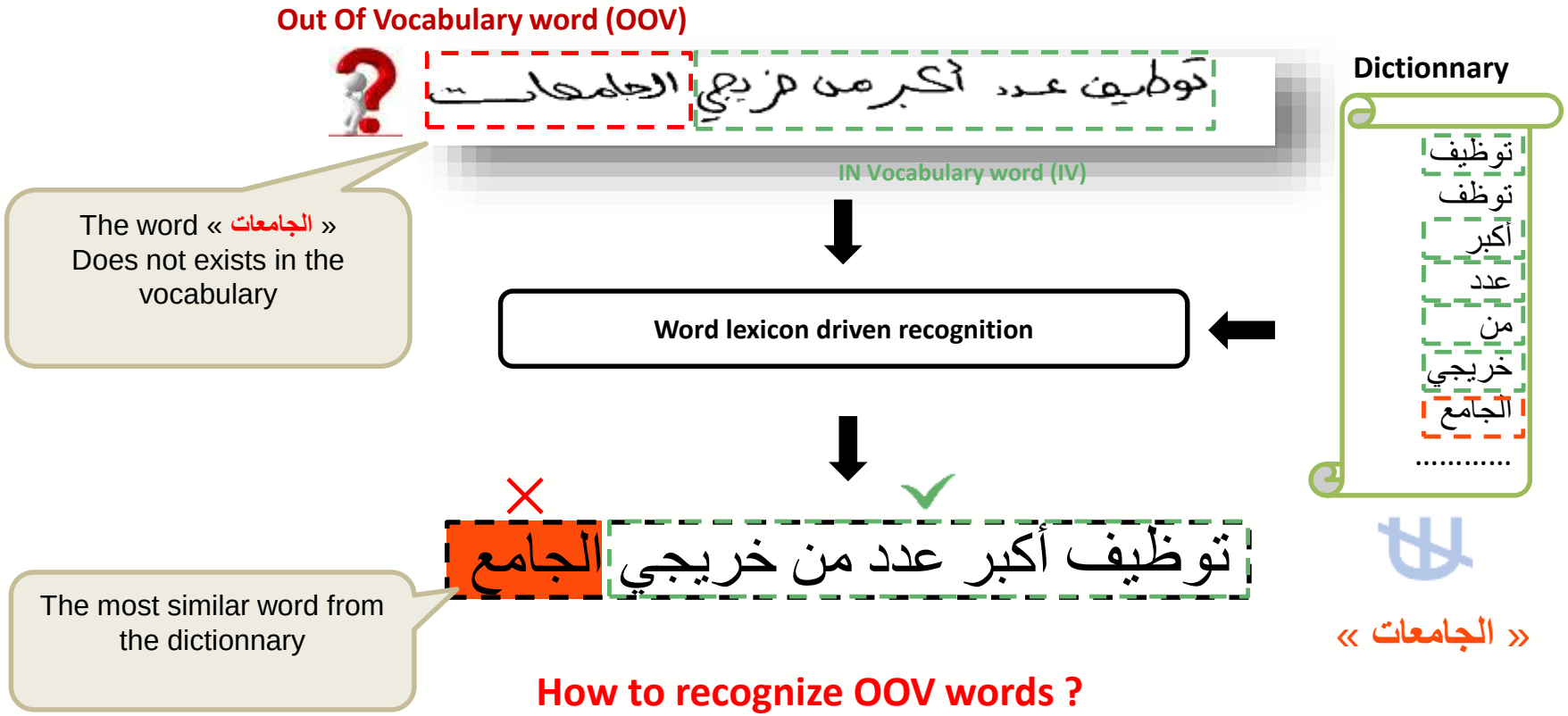
BUT…

Recognition still depends of linguistic resources

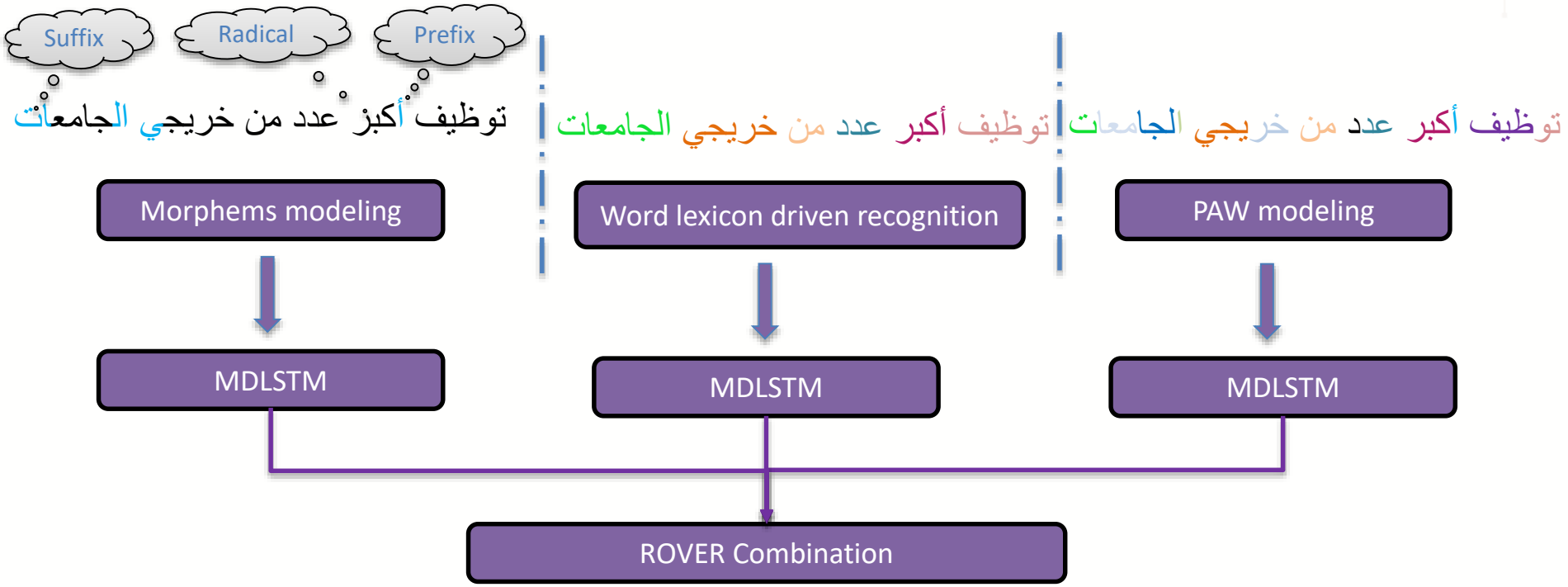
How to handle OOV words ?

Out-of-vocabulary words

- Handwriting recognition systems usually rely on static dictionaries
  - Small lexicon: Low coverage rate
  - Big lexicon: High computing time + high confusion between words
- Full coverage of these dictionaries is generally not achieved



Solution 1 : Combining different sub-word modeling

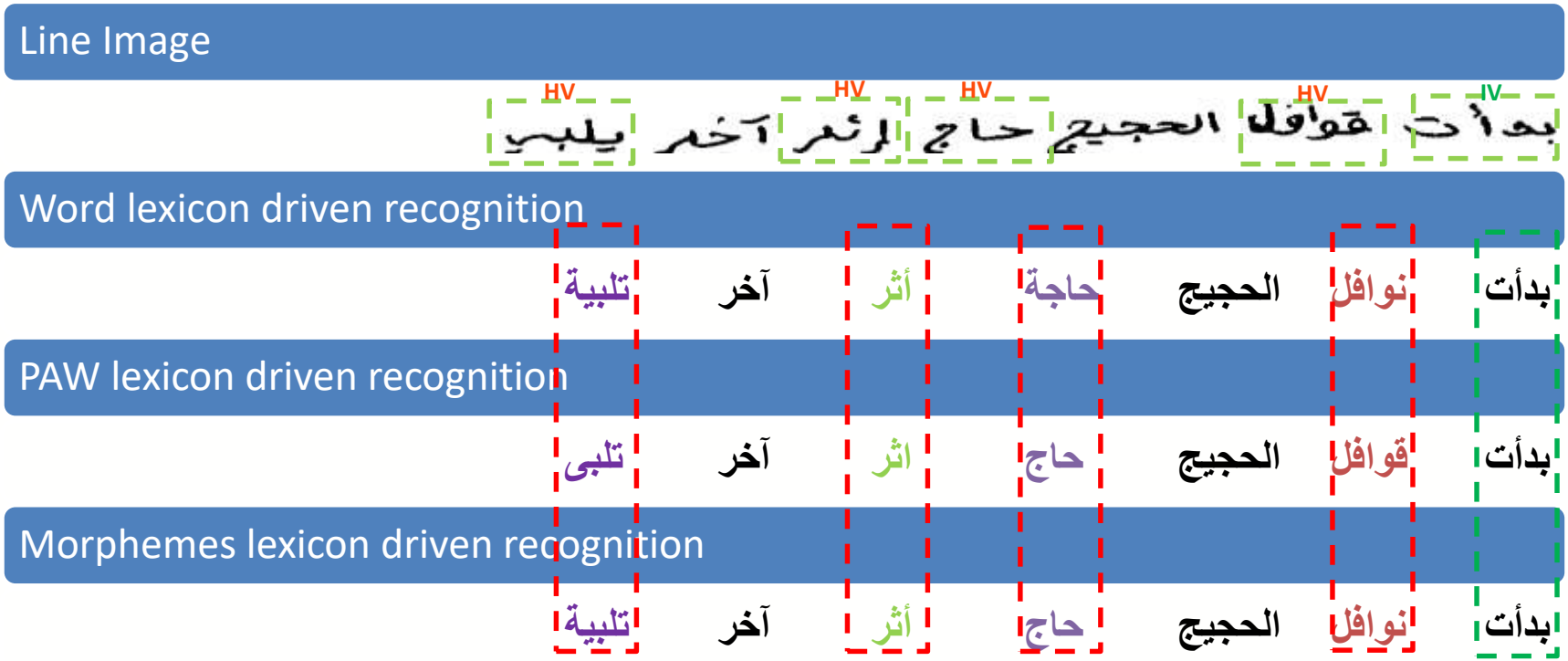


S. Khamekhem Jemni, **Y. Kessentini**, S. Kanoun, Improving Recurrent Neural Networks for Offline Arabic Handwriting Recognition by combining different Language Models, International Journal of Pattern Recognition and Artificial Intelligence. November 2019.

Solution 2 : OOV words detection and recovery using dynamic dictionnaries

Idea:

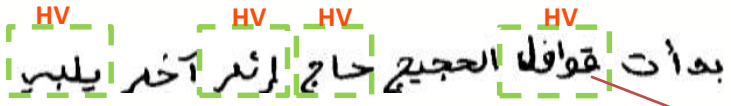
- Keep the IV words recognized by the word lexicon driven recognition
- Recover OOV word using dynamic lexicons built from large text corpora



OOV Detection stage

Solution 2 : OOV words detection and recovery using dynamic dictionnaries

OOV words Recovery stage



Identification of the Reference OOV

External WEB dictionnary

Lexicon Extension

Dynamique Lexicon driven recognition

Line text hypothesis

بدأت قوافل الحجيج حاج اثر آخر يلبي

Word lexicon driven recognition

نوافل 0.89

PAW lexicon driven recognition

قوافل 0.93

Morphemes Word lexicon driven recognition

نوافل 0.89

S. Khamekhem Jemni, Y. Kessentini, S. Kanoun, Out of Vocabulary Word Detection and Recovery in Arabic Handwritten Text Recognition, Pattern Recognition, Vol 93, pp. 507-520, 2019.

Results

- KHATT dataset
- Train : 9475 Test : 2007
- Lexicon : 18933 words, 11,46% OOV

Systems	WER
Hamdani et al. [2]	**26.80**
BenZeghiba et al. [1]	**30.9**
BenZeghiba [3]	**34.3**
Our approach	**20.83**

HIGH PERFORMANCE WITH DL

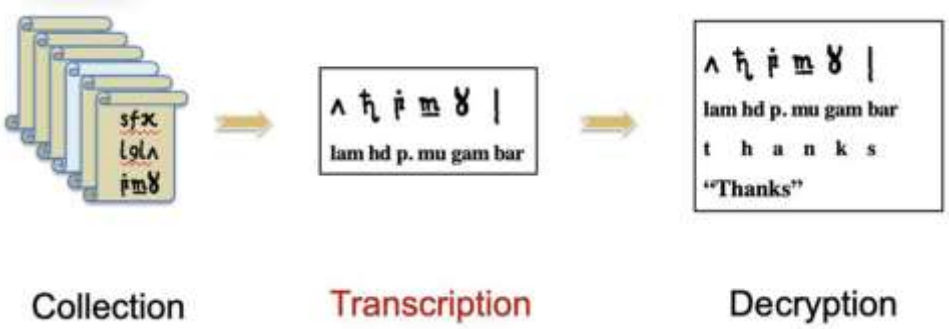
BUT…

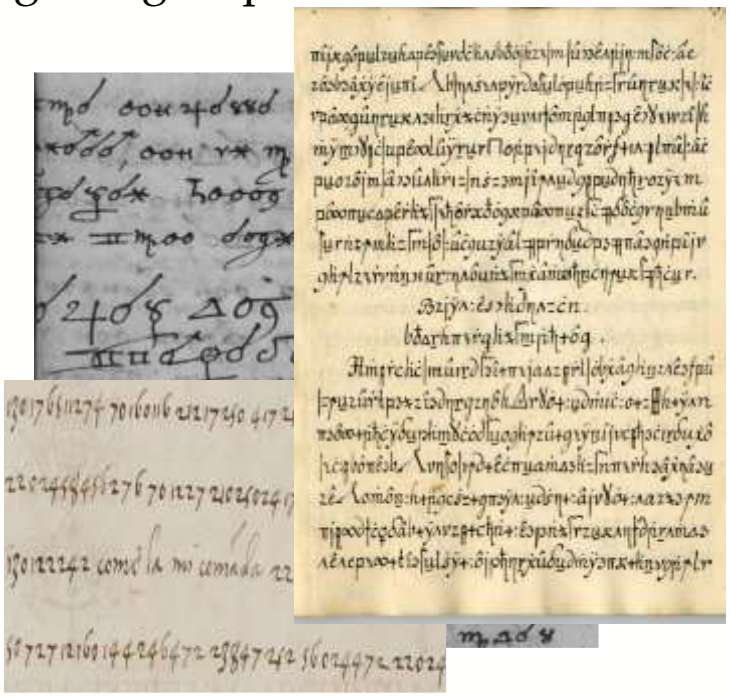
What about cases where you don't have labeled data?

Recognition of ciphered manuscript

- Around 1% of documents in archives contain encrypted text:
  - Diplomatic correspondence, secret societies/religious groups...

- DECODE/DECRYPT project:
  - To develop resources & tools for automatic decryption of enciphered documents from early modern times.

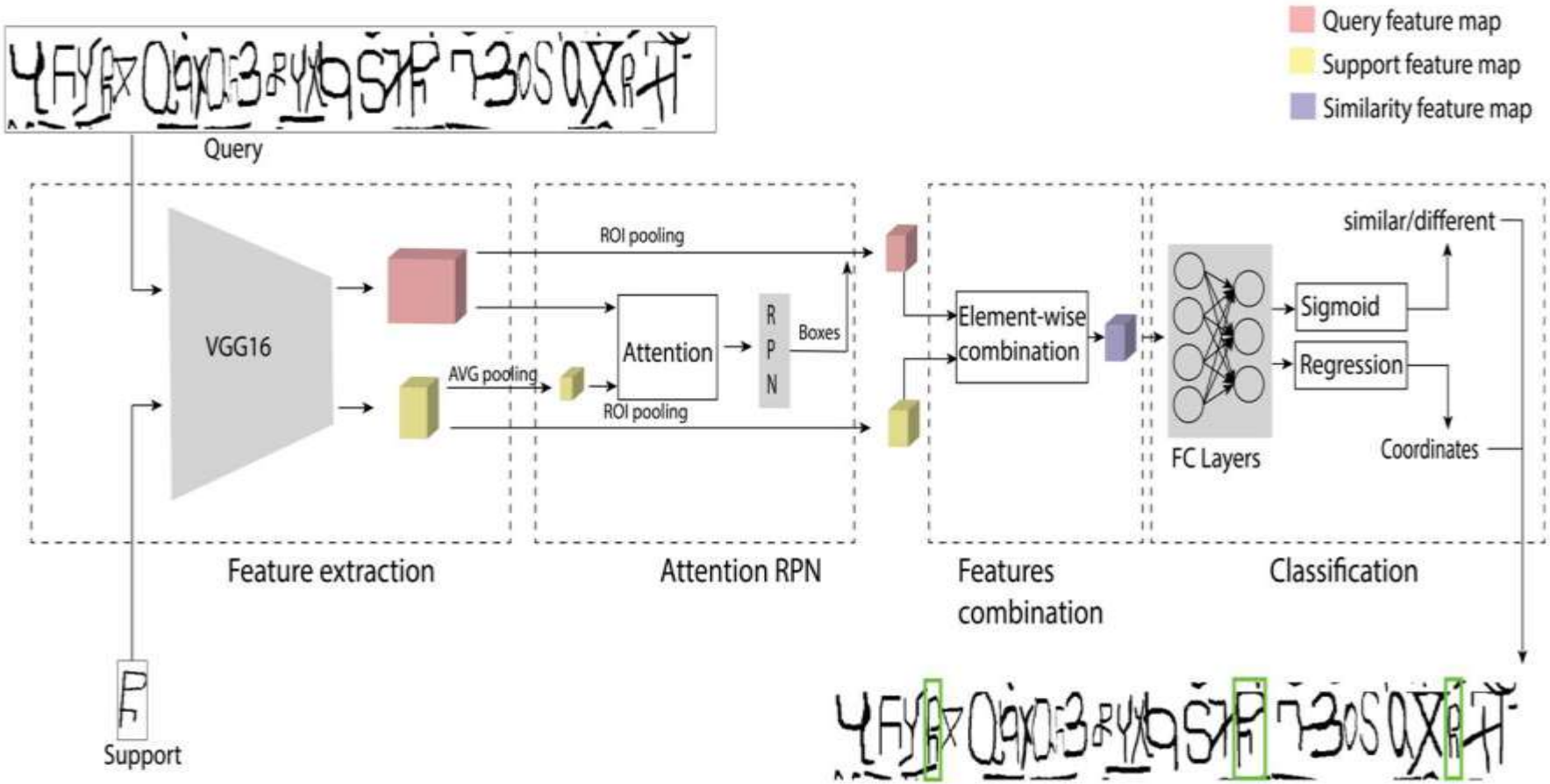




Problem : Very few (or none) labeled data to train

Few-shot Learning for Historical Ciphered Manuscript Recognition

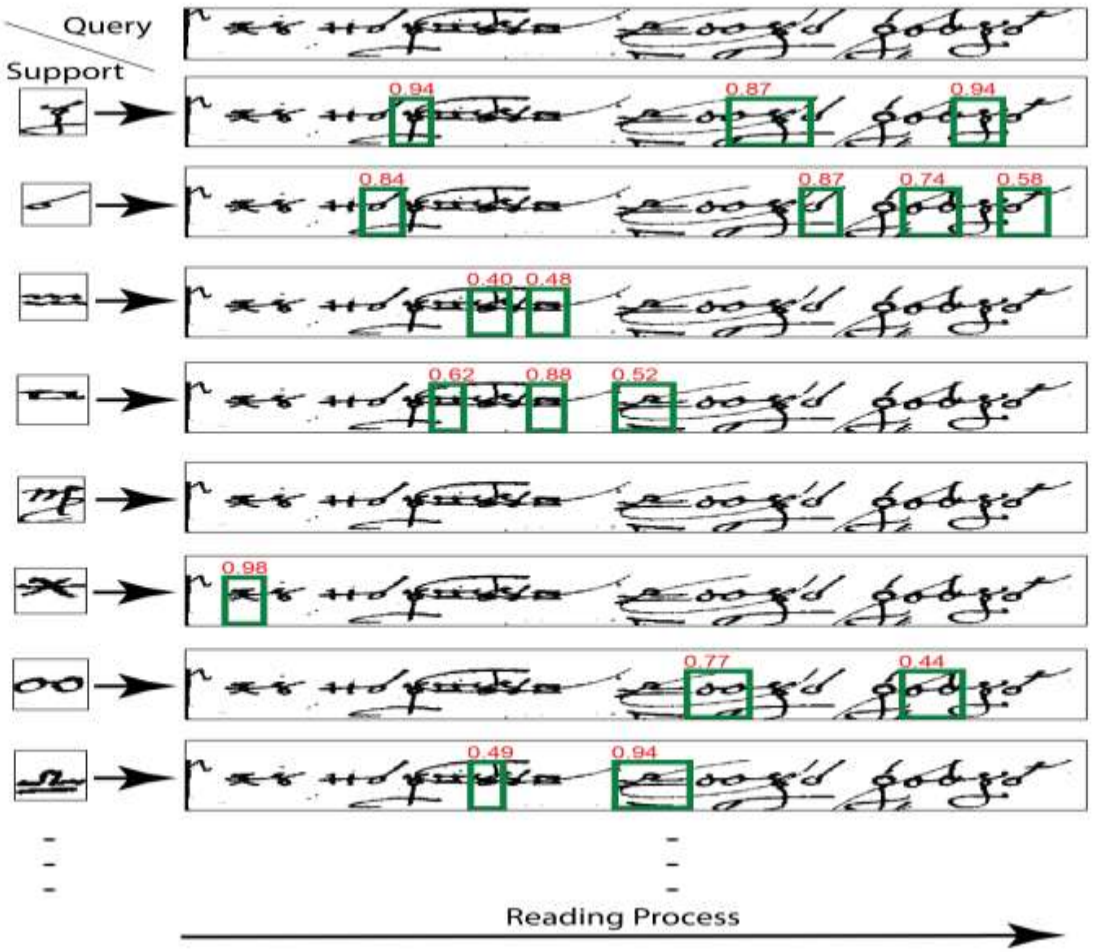
Symbol detection step



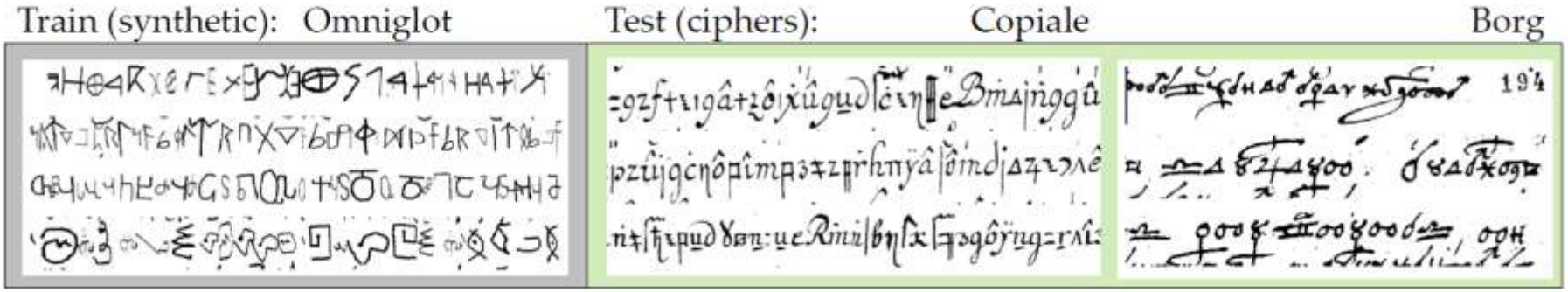
M.A.Souibgui, A.Fornés, Y.Kessentini, C.Tudor. A Few-shot Learning Approach for Historical Encoded Manuscript Recognition, International Conference on Pattern Recognition (ICPR), 2020.

Few-shot Learning for Historical Ciphered Manuscript Recognition

Text recognition step



Few-shot Learning for Historical Ciphered Manuscript Recognition



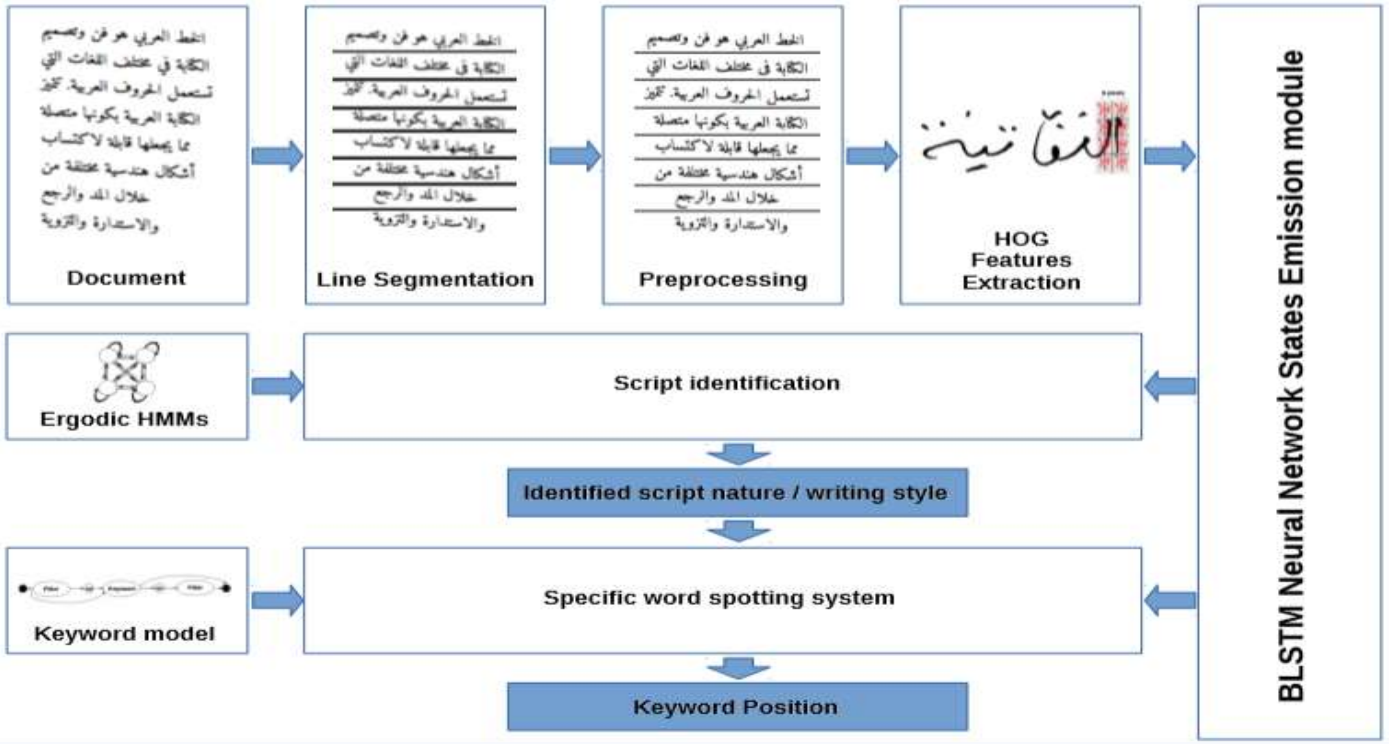
	Copiale			Borg		
Method	Training Pages	Shots	SER	Training pages	Shots	SER
LSTMs	25	-	0.11	7	-	0.55
	34	-	0.10	9	-	0.52
	42	-	0.07	11	-	0.45
Few-shot	2	1	0.10	2	1	0.17
		5	0.10		5	0.18

BUT…

DO WE NEED TO RECOGNIZE ALL THE DOCUMENT ?

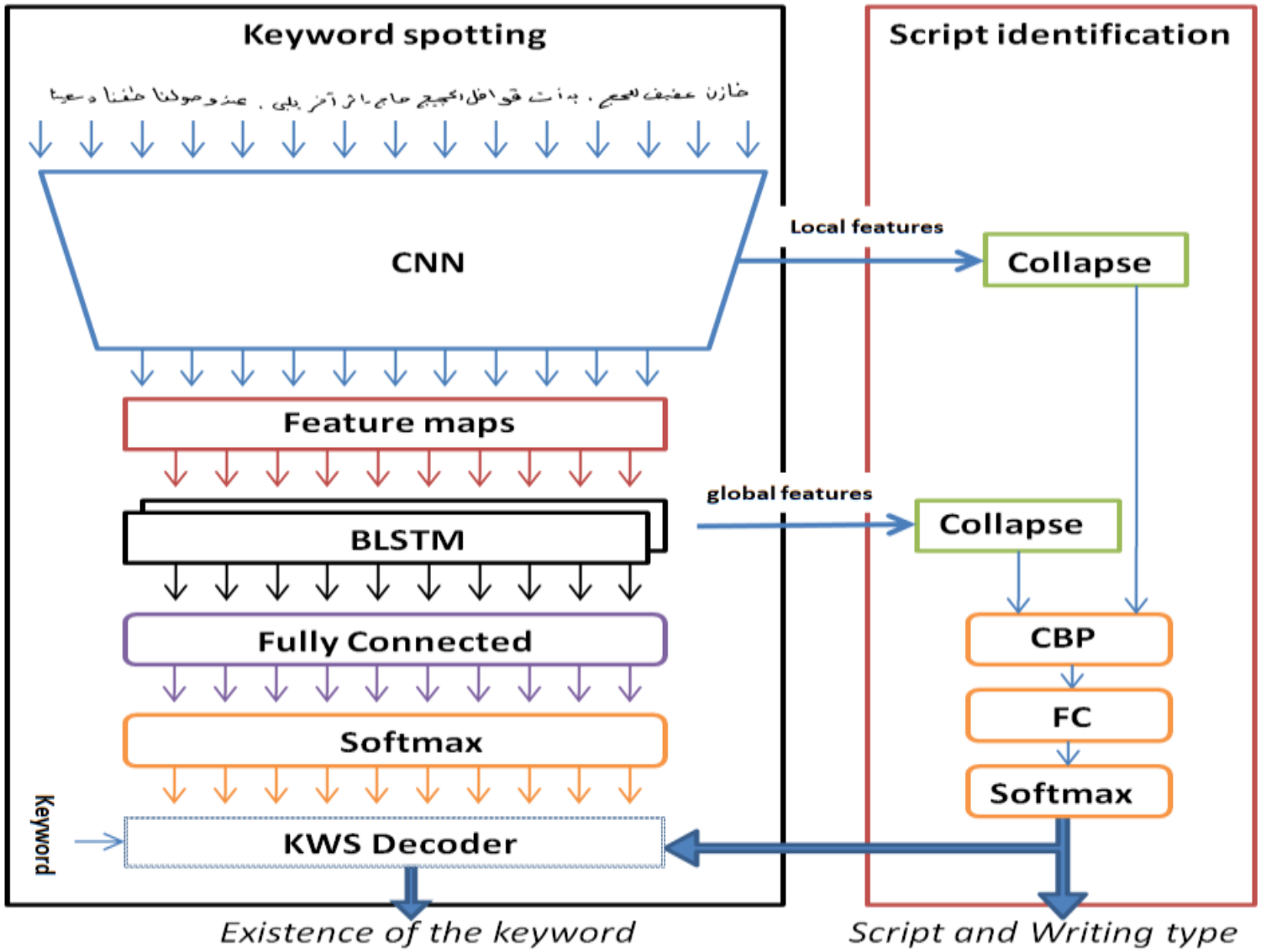
Keyword spotting

- Finding all instances of a query word that exist in a scanned document image, without fully recognizing it.
  - Document indexing and retrieval, routing, categorization…



A. Cheikhrouhou, **Y. Kessentini**, S. Kanoun, Hybrid HMM/BLSTM system for multi-script keyword spotting in printed and handwritten documents with identification stage, Neural Computing and Applications, vol. 32, pp. 9201–9215, 2020.

Multi-task learning



Results

Script identification results (Multi-task Learning)

Script	Handwritten French	Handwritten Arabic	Printed Arabic
Handwritten French	**100%**	0%	0%
Handwritten Arabic	0%	**99.9%**	0.1%
Printed Arabic	0%	0.11%	**99.89%**

Keyword spotting results (100 keywords MAP)

System	Handwritten French	Handwritten Arabic	Printed Arabic
Multi-Task	**89%**	**92.87%**	**98.29%**
BLSTM-HMM	70.12%	84.76%	82.15%
HMM	51.9%	49.1%	61.8%

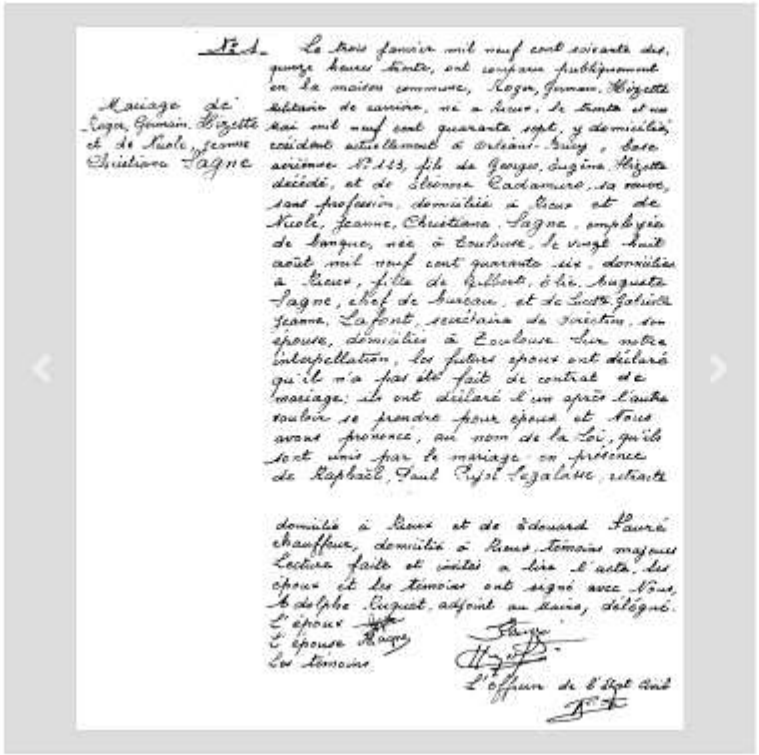
BUT…
IS RECOGNITION ENOUGH?

Recognition → Understanding

- READ project
  - **Information Extraction:** Transcription + Semantic Recognition (NER)
- Death, birth and marriage records

Files

1970M0001.tif



INFORMATION EPOUX

Nom Epoux: Eliette	Prénom Epoux: Roger German
Date Naissance: 31-5-1947	Ville: Rieux
Nom Père: Eugène	Prénom Père: Georges
Nom Mère: Laure,	Prénom Mère: Eléonore

INFORMATION EPOUSE

Nom Epouse: Lane	Prénom Epouse: Nicole Jeanne Christiane

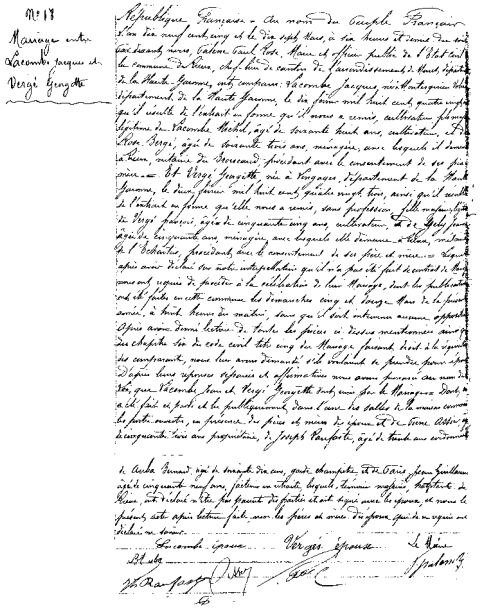
Conclusion

- Whenever I say that I am a researcher in Document Analysis & Recognition, people say :

I thought it was a solved problem !

Real data, realistic problem

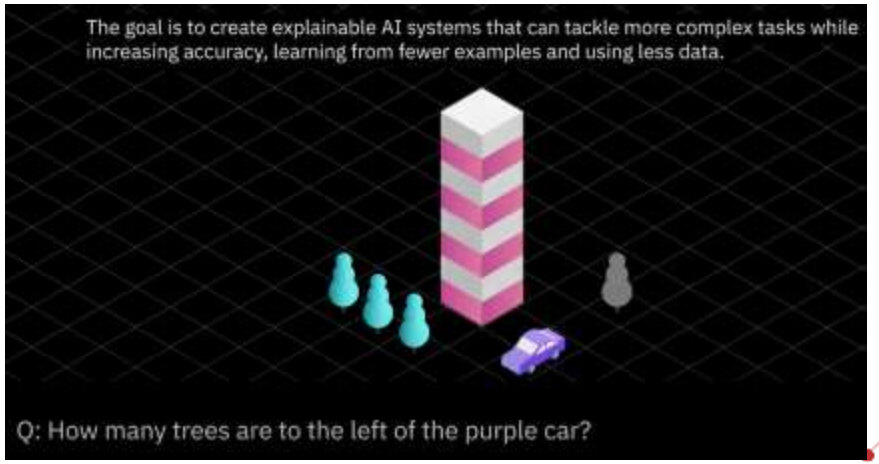
Real data, real problem



Very difficult to reach, even for industrial solution providers

The future

- Full-page Document recognition
- There will be a new approach to OCR
  - Seq2Seq, Transformers, Graph Neural Networks…
- **Labeled data** is an important limitation
  - Data generation, DomainAdaptation, Few/zero-shotLearning…
- Neuro-Symbolic AI
  - Represents the causal and compositional processes behind perceptual observations



Thank you!

Yousri Kessentini
Digital research center of Sfax

yousri.kessentini@crns.rnrt.tn
http://yousri.fr.cr