# 25<sup>th</sup> International Conference on Pattern Recognition (ICPR 2020)

**Combining Deep and Ad-hoc Solutions to Localize Text Lines in Ancient Arabic Document Images** 



Olfa Mechi<sup>1\*</sup>, Maroua Mehri<sup>1</sup>, Rolf Ingold<sup>2</sup> and Najoua Essoukri Ben Amara<sup>1</sup> <sup>1</sup>Laboratory of Advanced Technology and Intelligent Systems - ENISo, Tunisia

<sup>2</sup> DIVA, University of Fribourg, Switzerland

\* olfamechi@yahoo.fr





## **INTRODUCTION**

## **Context:**

- Advanced Partnership between the Laboratory of Technology and Intelligent Systems (LATIS, University of Sousse)<sup>[1]</sup> with the Tunisian national archives (ANT)<sup>[2]</sup> and our industry partner "Smart Information Trade"
- Historical documents transcription and indexing issues <sup>[3]</sup>



**Objective:** Text line localization in Arabic historical document images

**Challenges: Complexity** of particularities of historical the Arabic document images

#### **PROPOSED METHOD** Step 1 الافتراق وبيتا ومناوالسبك عندالالفا وبعين عليصد المزاج الزمين للعرابة الممتزجة فى داد الكون بقولم لا تعسل دوح جدد غير ولا تنت **Extracted text lines** ودج غجسد غرجا وبقولهمان الامثيا غاشراشكالها وتخالف اعتلاها $\rightarrow$ **U-Net architecture** ونبوكم ان علنا فى الاميدا التى ينها قرارة واحتمة لما يجعها من at X-height level القرابة الروحانية الستجنة فيواطنها وتقولون في يعوزهربا لمنكل Step 3 واحدمنها ظاهرا لدخر ويتك الغرابة يتوق بعضها الديعين وبليزم بعضها بععنا وباطن كل واحدمتها مستغرق مذور لادد ركه الحس ولا يكم بوجوده الاالعقل ولاسبيه لالخاخراج باطنه الخطاهم الإبامداده بأمتكا لدحت يصر المغلوب غالبا والباطن فلاصرا وبقولون اه ف Ascender and descender بواطن الاجساد ارواحا مستغنة هن قراميه الاجساد ولس لها وجود ظاهرجاسة منالحواس الااذ اامتدت باشكالها وإنابد لعلى بجود detection العقرا الذىهوا لماكم العدل اما أوماح الارداح فيد لعلى وجودها Step 2 الدالارداح تشابد الماوالشواع غلونها وصغاتها وتغالف فى الموامن الوجودة فيها وهوغيم وجودة غ الماء العراج وتلك الخاصية المادر وحاضة انشاتها فهاالتراكب المتعاقبة عليها فلغابعدخلق وإما المصروح اعنى ارواح الاجاد فدارعليها **Extracted** عم المرالتهبي فيهامع تسلطا النبران باورجم **Modified RLSA** $\rightarrow$ عليها وتلاحماجزاد المتكلس منها وعجزا الطوباس foreground contours المثاكلة لهاعن الوصول الى اهما قه الافتراق ونقاوم فارالسبك عندالالفا وبعين علىهذ االمزاج الزمين الافتراق ويقدوم نارالسبك عندالالفا ويعين على على المزاج الزمين للقرابة الممتزجة فى بدالكون بقولهم لاتقسل روح جسداغين ولانتبت للغرابة المتزجة فىدد الكون بقولهم لاتقسل روح جسداغرم كانشت روح غجسد غرجا وبقولهمان الامشا تنامل الشكالها وتخالين امنده وبقوكمان علمنا فالاسياالتى بينها قرادة وامتحة لما يجعهامن القزائية الروحافية المستجنة في واطنها ويقولون في رمون هرباط كل واحدمها ظاهوا لاخر وبتك الغرامة يتوق بعضها الربعض وبليزم

### RESULTS

#### **Table I: Performance evaluation of the proposed method**

#### on ancient Arabic and Latin document images

|              | Match | Miss | <b>False Alarm</b> | Split | Merge |
|--------------|-------|------|--------------------|-------|-------|
| RASM         | 1497  | 35   | 568                | 65    | 96    |
| ANT (Arabic) | 697   | 23   | 21                 | 12    | 0     |
| cBAD         | 6374  | 1106 | 865                | 439   | 39    |
| ANT (Latin)  | 838   | 11   | 5                  | 10    | 4     |

#### **Table II: Performance comparison of the five participating methods**

in RASM 2018 contest with the proposed method

|   | <b>Tesseract 3</b> | <b>Tesseract 4</b> | FRE11 | KFCN | RDI  | Proposed<br>method |
|---|--------------------|--------------------|-------|------|------|--------------------|
| Success rate<br>Text line correctly detected) | 28.8               | 44.2               | 43.2  | 67.7 | 81.6 | 88.42              |

والكاف والتان والتات والقار والظار بالتذوالذال والمار واللكر والعين والواء والخاء والمتان عواشتان Zoom اددت ان توامن عذه المقامات سيا فالتكن قراء فالق التىلدى اعدالمسومدوان واخد مادالد اورمعية لكون فيهاجن موافق الطبعة ذلك المقامي يظهر التلاء ذلك القاء على المادوالم المكاذا ودف

יון הליה לאומו שלאלי שלא אין אין

土地でないないといういんのないの

the state of the s

Bizing all the state of the second

And a state of the second seco

a alter a the sally way lines

Charles and a state of the second

المواليات فخرار المعاجد والمحاف المغال

and the telestate (12 and 12 and

SA ALLAS AN ACCOMENTS TO THE STORE

ويفاجده والمناف ومالا المسل



روح فجسد غيها وبقولهمان الاسيا تماشل اشكالها وتخالف اضادها وبعولهمان علمنا فالاستاالتى بينها وابة واستجة لما يجعهامن **Zoomed region** 

فالصرجاسة منالحواس الااذ اامتدت باشكالها وإماود إهلى العقلا لذىهوالحاكم العدل امآترواح الارداح فيد لطح وجودها أنالارداح تشابه الماءالضراح فلونها وصفاتها وتعالف فالغواس الموجردة فمها وهوغير وجوعة غالماء الغراح وتلك الخاصية أثادبر وحاضية أنشاتها فهاالتراكيب المتعاقسة عليه خلقابعنظق وإماالمروح اعف ارواح الاجساد فداعلم مسي ومرالتهبى فيهدامع تسلط النيران بلارحسي عليها وتلاحماجزاء المتكلس منها وعبزا لطويامت المثاكلة لماعن الوصول الى اعما قم

مضها بعضا وطاطن كل وإحدمتها مستغرق مذور لدديد كمالحس ولا

يمم بوجوده الاالعتل ولاسب لالاخلج بإطند الظاهره الابامداده

باشكا لدحق تعبر المناوير غالبا والباطن ظاهرا ويقولونه ان ف

بواطن الاجساد ارواحا مستجنة عن قراميه الاجساد وليس لها وجوم

**Fig 1: Scheme of the proposed method used for text line localization** in ancient handwritten Arabic document images

**Step 1:** Extraction of the X-height contour of each text line using U-Net architecture <sup>[3,4]</sup>

**Step 2:** Extraction of the foreground contour by applying a modified RLSA algorithm

**Step 3:** Determination of the ascender and descender components of each text line using topological structural analysis

## **EXPERIMENTAL CORPORA**





#### **Fig 4: Examples of some limitations of the proposed method**

## **CONCLUSIONS AND FUTURE WORK**

- ✓ Combining deep and ad-hoc solutions to localize text lines in ancient **Arabic document images**
- **Effectiveness and robustness of the proposed method in both Arabic** and Latin handwritten historical document images

**Fig 2: Examples of historical document images of the four evaluated datasets** 

## REFERENCES

[1] <u>http://www.latis-eniso.org/</u>

[2] <u>http://www.archives.nat.tn/</u>

[3] O. Mechi, M. Mehri, R. Ingold and N. Essoukri Ben Amara, "Text line segmentation in historical document images using an adaptive U-NET architecture," ICDAR, 2019. [4] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," MICCAI, 2015. [5] M. Kassis, A. Abdalhaleem, A. Droby, R. Alaasam and J. El-Sana, "VML-HD: The historical Arabic documents dataset for recognition systems," ASAR, 2017.

This work has been carried out as part of a research project (funded under the "19PEJC-08-02" grant agreement number) by the Tunisian Ministry of Higher Education and Scientific Research.

> Evaluate the proposed method on other public ancient handwritten Arabic document image datasets (e.g., Hadara and VML-HD)<sup>[5]</sup> > Propose a deep system able to transcribe printed and handwritten text lines from Arabic and Latin historical document images