Papy-S-Net : A Siamese Network to match papyrus fragments

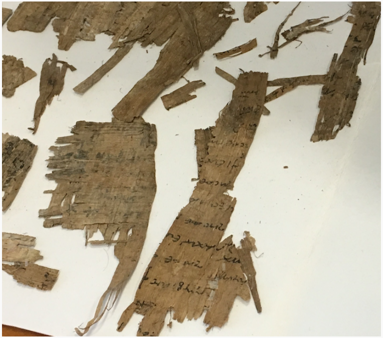
HIP 2019 Workshop, ICDAR, Sydney

Antoine Pirrone, Marie Beurton-Aimar, Nicholas Journet

September 20, 2019

Context

- *GESHAEM* Project (Archeological Project)[1]
- Digitalize and study the content of papyri

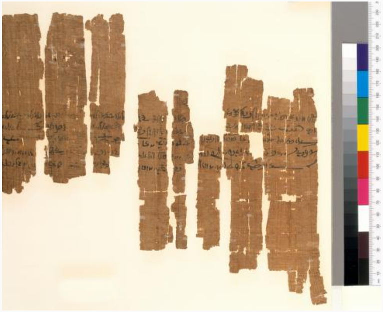


[1] This research has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme under grant agreement No 758907 and is part of the *GESHAEM* project.

Papyrologist work

Resolving a complex puzzle:[2]

- Laborious and time consuming task
- Specific field of document analysis relatively unstudied
- Helping the papyrologists with Image Processing

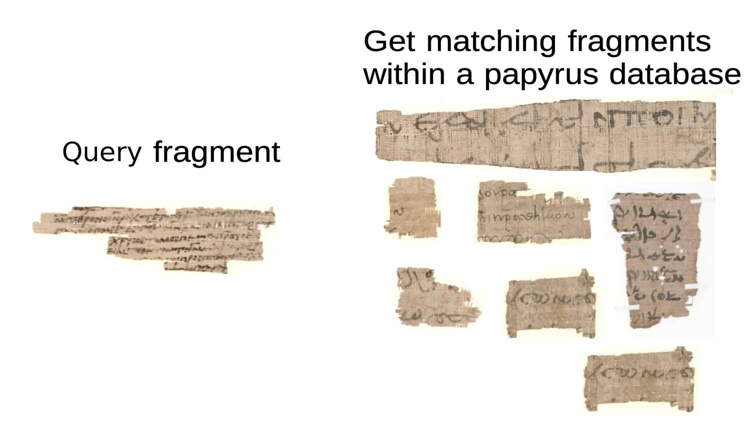


- 1 papyrus
- 12 fragments
- Had to be retrieved amongst several hundreds of fragments

[2] image from https://quod.lib.umich.edu/a/apis

Papyrologist work

First, sorting the pieces

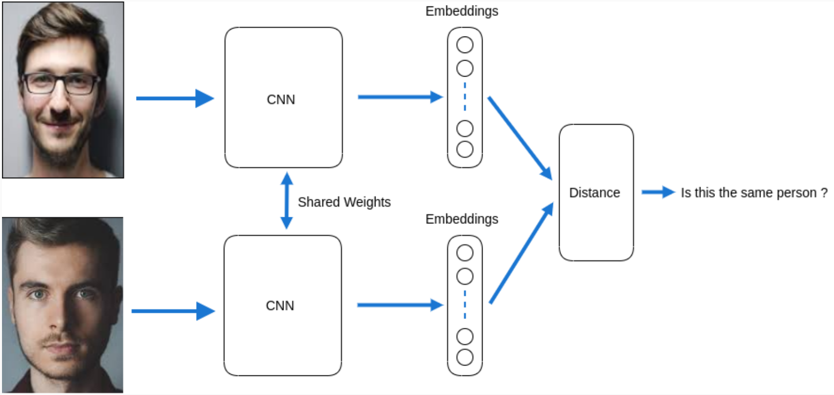


Query fragment

Get matching fragments within a papyrus database

A supervised learning approach

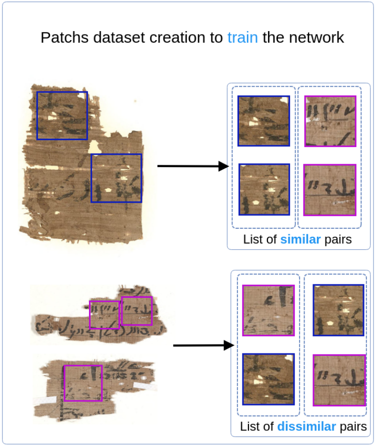
Training a Deep Siamese Network to know if two fragments are coming from the same papyrus



A supervised learning approach

Training a Deep Siamese Network to know if two fragments are coming from the same papyrus

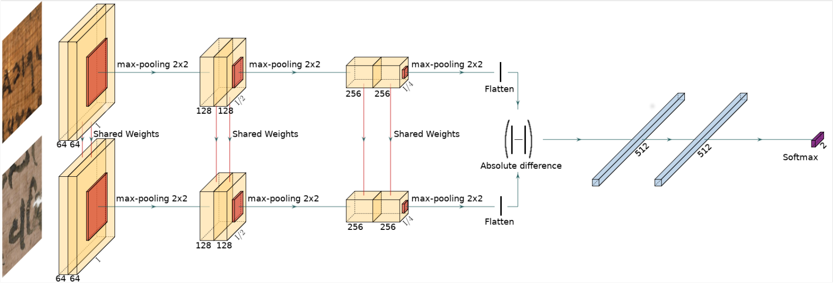
- Similar and dissimilar pairs to train the network
- Patch based approach



Papy-S-Net

A Siamese Deep Convolutional Neural Network[3]

- Fragment similarity → to belong to the same papyrus



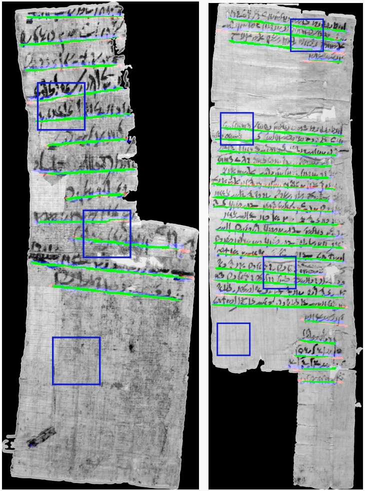
[3] Code available upon request

Impact of patch extraction method

Extracting patches:

- With text
- Without text
- Randomly

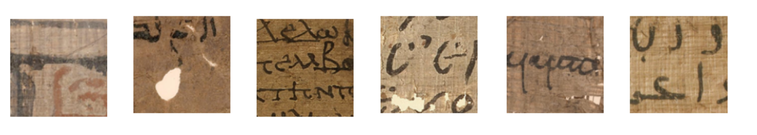
- Baseline segmentation to find where the text is
- All patches are the same size



Learning process

Our Dataset :

- 500 fragments [4] :
  - -600 to +400 BCE
  - In arabic, coptic, demotic, grec, hebrew, hieratic and latin
- 12.000 extracted patches for each method
- Train : 72%, Validation : 18%, Test : 10%

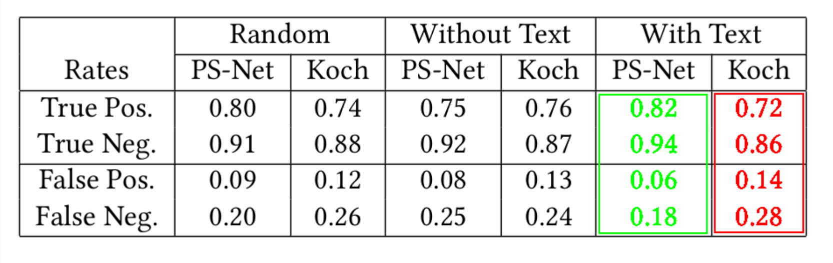


[4] coming from https://quod.lib.umich.edu/a/apis Accessed: June 04, 2019

Testing Papy-S-Net - on 10% of the dataset

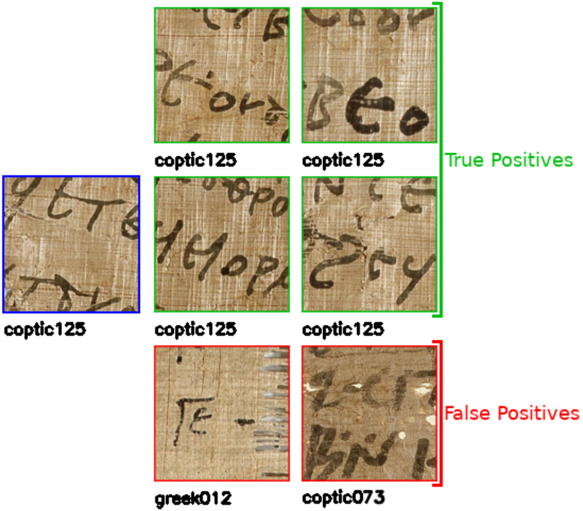
Results :

- Comparison with Koch et al.'s architecture (Koch et al. 2015)
- Best results with **Papy-S-Net** on patches *With text*

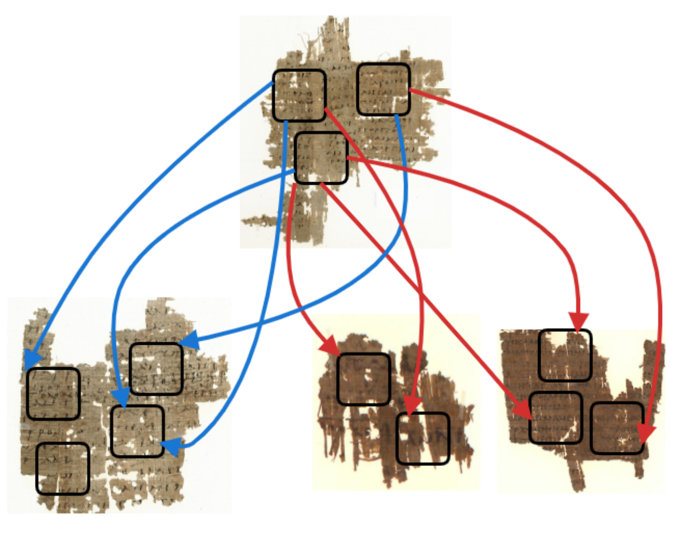


Rates	Random		Without Text		With Text	
	PS-Net	Koch	PS-Net	Koch	PS-Net	Koch
True Pos.	0.80	0.74	0.75	0.76	0.82	0.72
True Neg.	0.91	0.88	0.92	0.87	0.94	0.86
False Pos.	0.09	0.12	0.08	0.13	0.06	0.14
False Neg.	0.20	0.26	0.25	0.24	0.18	0.28

Examples of matchings

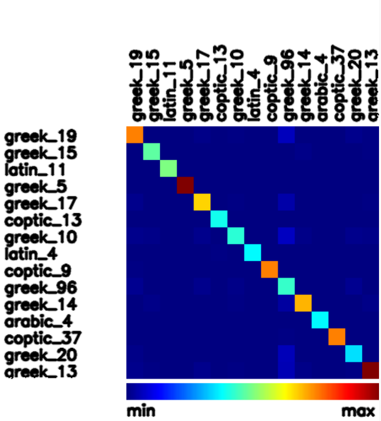


Testing a real use case



Testing a real use case

About 30 fragments from 15 papyri to reconstruct



- 89% True Positives
- 23% False Positives
- 77% True Negatives
- 11% False Negatives

Conclusion and Current works

Conclusion

- Proposed a Siamese architecture adapted to papyrus fragments matching.
- Obtained 89% of true positives on a real use case test.
- A good first step towards more advanced works.

Current works

- Building bigger database ($\sim$ 15.000 fragments, $\sim$ 1000 reconstructed papyri, ground truth).
- Applying on other databases.
- Experiments with Triplet Networks (Hoffer and Ailon, 2015).

References i

- P. Butler, P. Chakraborty, and N. Ramakrishan.
  The deshredder: A visual analytic approach to reconstructing shredded documents.
  In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 113–122. IEEE, 2012.

- T. Grüning, G. Leifert, T. Strauß, and R. Labahn.
  A two-stage method for text line detection in historical documents.
  arXiv preprint arXiv:1802.03345, 2018.

References ii

E. Hoffer and N. Ailon.
Deep metric learning using triplet network.
In A. Feragen, M. Pelillo, and M. Loog, editors, *Similarity-Based Pattern Recognition*, pages 84–92, Cham, 2015. Springer International Publishing.

G. R. Koch.
Siamese neural networks for one-shot image recognition.
2015.

G. Levi, P. Nisnevich, A. Ben-Shalom, N. Dershowitz, and L. Wolf.
A method for segmentation, matching and alignment of dead sea scrolls.

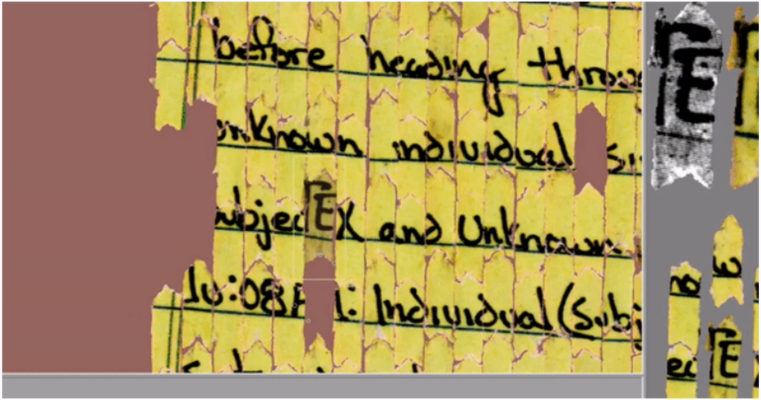
References iii

In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 208–217. IEEE, 2018.

Z. Zhong, W. Pan, L. Jin, H. Mouchre, and C. Viard-Gaudin.
Spottingnet: Learning the similarity of word images with convolutional neural network for word spotting in handwritten historical documents.
In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 295–300, Oct 2016.

Related Work

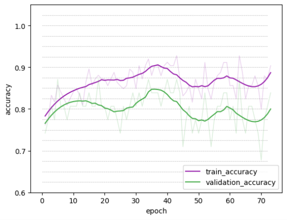
- Mainly methods for recovering shredded documents (Butler et al. 2012)



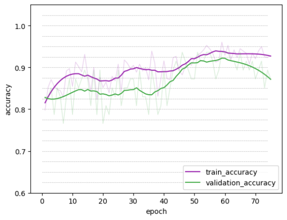
before heading throu
nknown individual S
ubjectEX and Unknow
16:08 1: Individual (Subj

- Optimization problem (text/shape/color continuity)
- Crowd sourcing problem

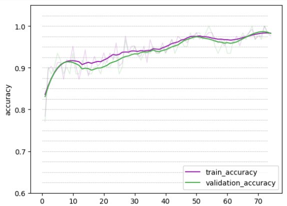
Learning process - (Training/Validation) on 90% of the dataset



1. Patches containing only texture



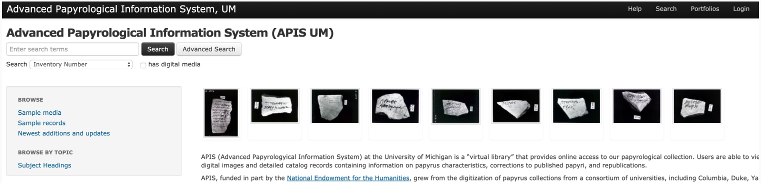
2. Random patches



3. Patches all containing text

A common objective for many projects

- *Michigan Collection* : 26.000 papyri

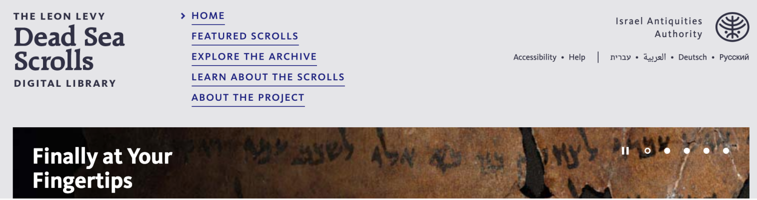


Advanced Papyrological Information System (APIS UM)

APIS (Advanced Papyrologyical Information System) at the University of Michigan is a “virtual library” that provides online access to our papyrological collection. Users are able to vie digital images and detailed catalog records containing information on papyrus characteristics, corrections to published papyri, and republications.

APIS, funded in part by the National Endowment for the Humanities, grew from the digitization of papyrus collections from a consortium of universities, including Columbia, Duke, Ya

- *Dead Sea Scrolls Collection* : 2000 papyri



- *GESHAEM* project (4 years) : 500 fragments to reconstruct

Related Work

For Papyrus

- Improve the digitalization process
- Identify duplicated fragments (Levi et al. 2018)