

Arabic Manuscript Author Verification Using Deep Convolutional Networks

Andrei Boiarov*, Alexander Senov†, Alexander Knysh‡ and Dmitry Shalymov§

*†§ Faculty of Mathematics and Mechanics

Saint Petersburg State University

Saint Petersburg, Russia

Email: *a.boiarov@spbu.ru, †alexander.senov@gmail.com, §dmitry.shalymov@gmail.com

‡Department of Near Eastern Studies

University of Michigan

Ann Arbor, Michigan 48104-1608, USA

Email: alknysh@umich.edu

Abstract—Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetur.

I. INTRODUCTION

[Replace dummy text with normal] Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetur at, consectetur sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

Morbi luctus, wisi viverra faucibus pretium, nibh est placerat odio, nec commodo wisi enim eget quam. Quisque libero justo, consectetur a, feugiat vitae, porttitor eu, libero. Suspendisse sed mauris vitae elit sollicitudin malesuada. Maecenas ultricies eros sit amet ante. Ut venenatis velit. Maecenas sed mi eget dui varius euismod. Phasellus aliquet volutpat odio. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque sit amet pede ac sem eleifend consectetur. Nullam elementum, urna vel imperdiet sodales, elit ipsum pharetra ligula, ac pretium ante justo a nulla. Curabitur tristique arcu eu metus. Vestibulum lectus. Proin mauris. Proin eu nunc eu urna hendrerit faucibus. Aliquam auctor, pede consequat laoreet varius, eros tellus scelerisque quam, pellentesque hendrerit ipsum dolor sed augue. Nulla nec lacus.

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetur odio sem sed wisi.

This research is supported by Saint-Petersburg State University grant 6.37.181.2014. The authors express their deep gratitude to Mrs. Evyn Kropf of the Hatcher Graduate Library who kindly facilitated access to the University of Michigan library resources.

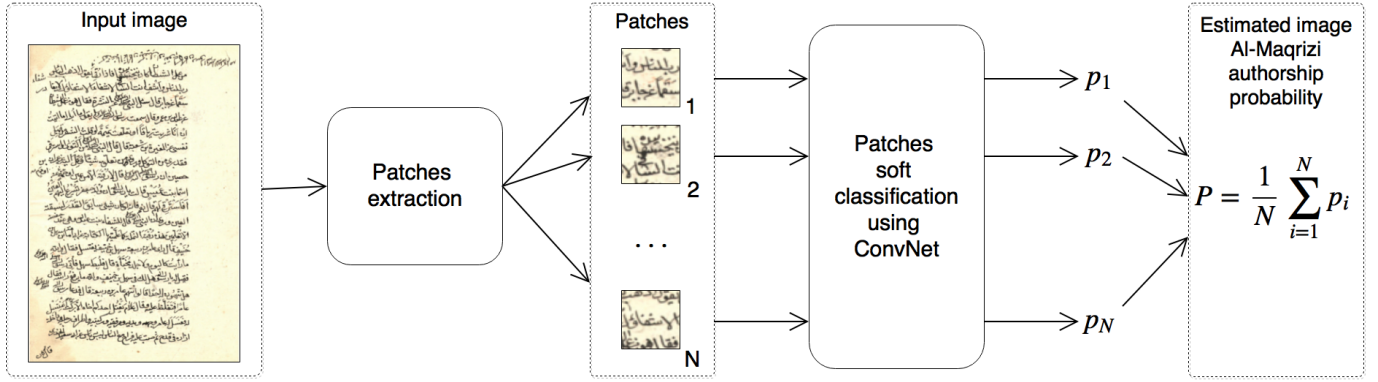


Fig. 1. Al-Maqrizi authorship soft classification pipeline

II. THE DATA

[Replace dummy text with normal] Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetur eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris. Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdiet justo nec dolor.

Etiam euismod. Fusce facilisis lacinia dui. Suspendisse potenti. In mi erat, cursus id, nonummy sed, ullamcorper eget, sapien. Praesent pretium, magna in eleifend egestas, pede pede pretium lorem, quis consectetur tortor sapien facilisis magna. Mauris quis magna varius nulla scelerisque imperdiet. Aliquam non quam. Aliquam porttitor quam a lacus. Praesent vel arcu ut tortor cursus volutpat. In vitae pede quis diam bibendum placerat. Fusce elementum convallis neque. Sed dolor orci, scelerisque ac, dapibus nec, ultricies ut, mi. Duis nec dui quis leo sagittis commodo.

Aliquam lectus. Vivamus leo. Quisque ornare tellus ullamcorper nulla. Mauris porttitor pharetra tortor. Sed fringilla justo sed mauris. Mauris tellus. Sed non leo. Nullam elementum, magna in cursus sodales, augue est scelerisque sapien, venenatis congue nulla arcu et pede. Ut suscipit enim vel sapien. Donec congue. Maecenas urna mi, suscipit in, placerat ut, vestibulum ut, massa. Fusce ultrices nulla et nisl.

Etiam ac leo a risus tristique nonummy. Donec dignissim tincidunt nulla. Vestibulum rhoncus molestie odio. Sed lobortis, justo et pretium lobortis, mauris turpis condimentum augue, nec ultricies nibh arcu pretium enim. Nunc purus neque, placerat id, imperdiet sed, pellentesque nec, nisl. Vestibulum imperdiet neque non sem accumsan laoreet. In hac habitasse platea dictumst. Etiam condimentum facilisis libero. Suspendisse in elit quis nisl aliquam dapibus. Pellentesque auctor sapien. Sed egestas sapien nec lectus. Pellentesque vel dui vel neque bibendum viverra. Aliquam porttitor nisl nec pede. Proin mattis libero vel turpis. Donec rutrum mauris et libero. Proin euismod porta felis. Nam lobortis, metus quis elementum commodo, nunc lectus elementum mauris, eget vulputate ligula tellus eu neque. Vivamus eu dolor.

III. THE METHOD

We consider author verification problem as a binary classification problem: Al-Maqrizi class denoted as 1 and non-Al-Maqrizi class denoted as 0. In this context our goal is to build a classification pipeline able predict the probability (*soft* classification) that given image belongs to the 1 (Al-Maqrizi) class. The entire Al-Maqrizi authorship classification pipeline illustrated at figure 1 consists of the following steps:

- 1) Image preprocessing.
- 2) Extracting patches from candidate image.
- 3) Patches soft classification using ConvNet .
- 4) Averaging predicted patches probabilities to produce overall candidate image Al-Maqrizi authorship probability.

Each of this steps are thoroughly described in the following sections.

[Describe image preprocessing]

A. Patches extraction

The patches extraction method generates a set of sub-images called patches from given image. The basic idea is that patch should represent small but yet meaningful part of image for the main purpose - authorship verification. We use two alternative methods for patches extraction described in following subsections.

1) *Sliding window based method*: This method splits given image by a grid of fixed cell size. Each cell further used as a patch. Figure ?? [add figure] illustrates the idea.

2) *Connected components based method*: This method uses following routine for patches extraction

- 1) Input image binarization using Otsu's filter [reference to Otsu filter paper].
- 2) Connected components extraction from binarized image using algorithm from [reference to connected components paper]. c
- 3) Too small, too big and too stretched connected components filtering using several empirical rules [specific rules description].
- 4) Outlier connected components filtering using DBSCAN [reference to DBSCAN paper] clustering algorithm.

- 5) Using remaining connected components bounding boxes as patches.

Example of connected components based patches show on figure ?? **[add figure]**.

It could be seen, that connected components based patches usually consist of one or few letters thus providing high robustness for different image scale and size in contrast to fixed-size sliding window patches. However, fixed-sliding patches contain much more information: several symbols from several lines, - a very important feature for the authorship verification.

[Fill this section]

IV. RESULTS AND DISCUSSION

[Fill this section]

V. CONCLUSION

[Fill this section]

REFERENCES

- [1] M. Bulacu, L. Schomaker, A. Brink "Text-independent writer identification and verification on offline arabic handwriting," in *Proc. 9th International Conference on Document Analysis and Recognition, ICDAR*, Curitiba, 2007, pp. 769–773.
- [2] D. Fecker, A. Asi, W. Pantke, V. Mrgner, J. El-Sana, T. Fingscheidt "Document Writer Analysis with Rejection for Historical Arabic Manuscripts," in *Proc. 14th nternational Conference on Frontiers in Handwriting Recognition, ICFHR*, Crete, 2014, pp. 743–748.
- [3] Y. Lecun, Y. Bengio, G. Hinton, "Deep learning," *Nature*, no. 521, pp. 436–444, May. 2015.
- [4] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, "Gradient-based learning applied to document recognition," in *Proc. of the IEEE*, 1998, pp. 2278–2324.
- [5] A. Krizhevsky, I. Sutskever, G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, vol. 25, 2012, pp. 1097–1105.
- [6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, "Going deeper with convolutions," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 2015, pp. 1–9.
- [7] O. Granichin, V. Volkovich, D. Toledano-Kitai, *Randomized Algorithms in Automatic Control and Data Mining*. Springer-Verlag: Heidelberg, New York, Dordrecht, London, 2015, 251 p.