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
St. Petersburg State University Laboratory of Analysis and Modeling of Social Processes
Ann Arbor, Michigan, 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 consectetuer.**

I. Introduction

The present study was motivated by the recent discovery by Dr. Noah Gardiner of the holograph (autograph) copy of the third volume of al-Maqrizi's famous Description of Egypt in the Library of the University of Michigan (Michigan Islamic MS 605) [1]. The full title of the manuscript is "al-Mawa'iz wa-al-i'tibar fi dhikr al-khitat wa-al-athar" ("The Book of Admonitions and Lessons in the Catalogue of Territorial Divisions and Historical Monuments"; usually cited simply as al-Khitat). The manuscript was copied well after 818 A.H. (1415 C.E.) and was finished shortly after 831 (1427) by the celebrated Egyptian historian Taqi al-Din Ahmad Ibn 'Ali al-Maqrizi (d. 845/1442). It is the only known Maqrizi holograph (autograph) in the Americas. A number of elements in the codex, including the apparent age of the paper, lacunae in the text where the dates of certain events had not been filled in, and a number of marginal addenda and sewn-in inserts containing text found in the printed editions, led Dr. Gardiner to suspect that it might be a draft copy of the work. He visually collated the predominant hand of the codex and the inserts with some published images of al-Maqrizis hand and felt that a match was highly likely. He then sent images of the codex

to Prof. Frederic Bauden of the University of Liege, the author of numerous articles on al-Maqrizi autographs. Prof. Bauden confirmed that the codex was indeed copied by al-Maqrizi himself, and was thus a holograph (autograph). He identified it as the fair copy (the authors final version) of the third volume of al-Khitat, and thus the only fair copy of any volume of al-Khitat to have been found.

Given the importance of this discovery for the history of science (al-Maqrizi's Khitat is one of the earliest descriptions of the topography of Cairo and ancient Egyptian monuments in its environs as well as Alexandria), a cross-disciplinary team of researchers affiliated with the St. Petersburg State University (Laboratory of Analysis and Modeling of Social Processes) decided to verify Dr. Gardiners and Prof. Baudens findings by using method based on deep convolutional networks.

Previously used methods for author verification of arabic manuscript and for related fields were focused on developing of various types of features that can be obtained from a manuscript picture [2], [3]. In this paper we present a novel approach based on learning of deep hierarchical structure of features from the row image. This method belongs to the class of deep learning algorithms [4] and uses convolutional network [5] for feature extraction and prediction. Deep convolution networks since 2012 year [6] became the state of the art in many areas of computer vision: objects recognition, face identification, optical character recognition, object detection etc [4].

The paper is organized as follows. In the next section is given a description of data set used in experiments. Section III contains presentation of our methodology of extracting patches from image, training deep convolution networks and making decision of manuscript author. In section IV we present results of our experiments. At the end, conclusion is given.

II. Data Set

Solving of the problem of al-Maqrizi authorship verification requires training and validation data sets. As a one unique
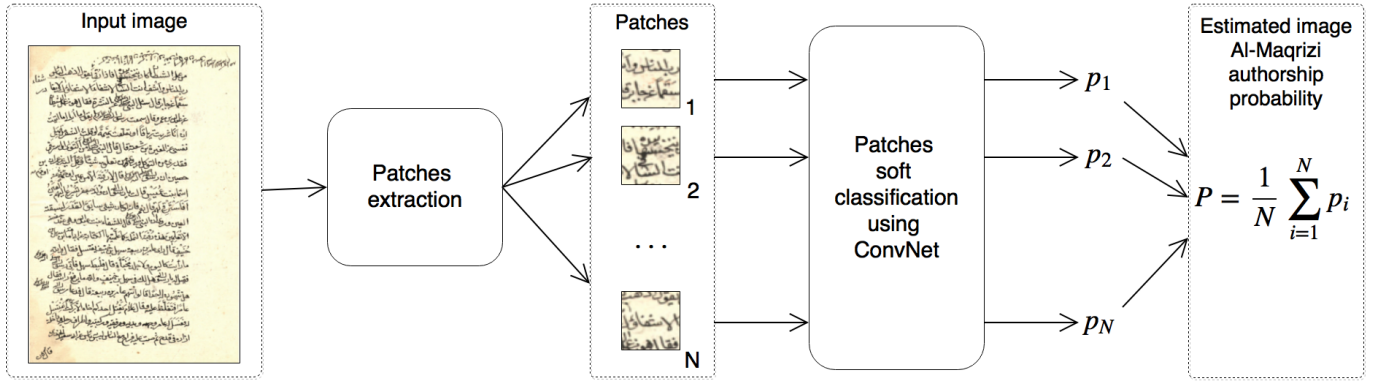
Figure 1. Al-Maqrizi authorship soft classification pipeline

data element we consider single page of manuscript. Data set composes of a set of consistent parts from manuscripts of two kinds:

- Two sets of verified al-Maqrizi's holographs from codex of Prof. Bauden is taken as a positive examples.
- Eight manuscripts not written by al-Maqrizi's hand from the University of Michigan Hatcher Library (Special Collections) are considered as a negative examples. This manuscripts are selected in such way that the date and place of writing each of them is close to al-Maqrizi's Khitat: 14th and 15th centuries, Egypt and Syria.

For robustness of learning process obtained data is divided into training and validation sets:

- Training set: 1 al-Maqrizi's manuscript consisting of 26 pages and 5 not al-Maqrizi's manuscripts each of which consists of 7 pages.
- Validation set: 1 al-Maqrizi's manuscript consisting of 14 pages and 3 not al-Maqrizi's manuscripts each of which consists of 7 pages. Authors of this 3 manuscripts differ from authors of 5 negative examples in training set.

It is important to note that we split our data set by the factor of a manuscript author. Thus when algorithm is learning on training set, it can not see authors in validation set. In that way we achieve robustness our method in terms of manuscripts because input of method is a whole manuscript. Number of not al-Maqrizi's documents is chosen in such way that training and validation set is nearly balanced.

Main interest of this paper is author verification of al-Khitat manuscript consisting of 32 pages.

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

The join research on al-Maqrizis Description of Egypt undertaken by a historian-philologist and three mathematicians from St. Petersburg State University is a unique experiment in working across disciplinary boundaries to achieve a common goal. Its results bode well for the future by opening new horizons for scholars of Oriental manuscripts who often desperately lack resources (other than their own eyes and intuition) to verify the provenance and authorship of the manuscript material they are working with. Given the propensity of Muslim scribes and later writers to attribute manuscripts to important luminaries of the past (such as, e.g., al-Ghazali, d. 505/1111; Ibn al-Arabi, d. 638/1240, and others), the new methods of analyzing and verifying handwritten texts, which have been designed and tested by St. Petersburg mathematicians, are bound to become an important tool for their colleagues in the humanities and social sciences.

REFERENCES

[1] N. Gardiner, F. Bauden, "A Recently Discovered Holograph Fair Copy of al-Maqrzs al-Mawiz wa-al-itibr f dhikr al-khitat wa-al-thr (Michigan Islamic MS 605)," in *Journal of Islamic Manuscripts*, vol. 2, E. J. Brill, Leiden and Boston, 2011, pp. 123–131.

[2] 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.

[3] 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.

[4] Y. Lecun, Y. Bengio, G. Hinton, "Deep learning," *Nature*, no. 521, pp. 436–444, May. 2015.

[5] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, "Gradient-based learning applied to document recognition," in *Proc. of the IEEE*, 1998, pp. 2278–2324.

[6] 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.

[7] 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.

[8] 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.