

Convolutional Neural Networks For Arabic Manuscript Authorship Verification

Andrei Boiarov^{1,*}, Alexander Senov¹, Alexander Knysh² and Dmitry Shalymov¹

¹ Faculty of Mathematics and Mechanics, Saint Petersburg State University

Saint Petersburg, Russia

² Dept. of Near Eastern Studies, University of Michigan,
St. Petersburg State University Laboratory of Analysis and Modeling of Social Processes
Ann Arbor, Michigan, USA

* Corresponding author: andrei.boiarov@gmail.com

Abstract—In this paper consider the problem of an automatic authorship verification of manuscripts written in Arabic. This manuscripts are treated as an images. We propose the method based on accordingly trained deep convolutional neural network binary classifier which using as an inputs patches from image. We consider two types of patches extraction: sliding window and connected components. Obtained probabilities utilize for defining of the manuscript authorship. This method we use for author verification of the work of medieval Egyptian historian al-Maqrizi *al-Khitat* which volume was recently discovered. For training deep network was used a number of labeled data from al-Maqrizi and non-al-Maqrizi manuscripts. Our method demonstrate promising results in this problem and thus is a useful tool for historians and philologists working in the area of authorship verification.

Keywords—Deep learning; Author verification; Historical Arabic manuscript; Handwriting recognition; Convolutional networks; Image classification; Machine learning

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-Maqrizi's 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 author's 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 *al-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. Gardiner's and Prof. Bauden's 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 raw image. This method belongs to the class of the 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 section II a description of the data set used in experiments is given. Section III contains presentation of our methodology of patches from image extraction, training deep convolution networks and making decision of manuscript author. In section IV we present the results of our experiments, whereupon we provide a conclusion.

II. DATA SET

Solving the problem of al-Maqrizi authorship verification requires training and test data sets. We consider a single page of the manuscript as one data unit. As a one unique data element we consider single page of manuscript. The data set consists of two sets of manuscript materials:

- two holographs of al-Maqrizi verified by Prof. Bauden will serve as a positive example;

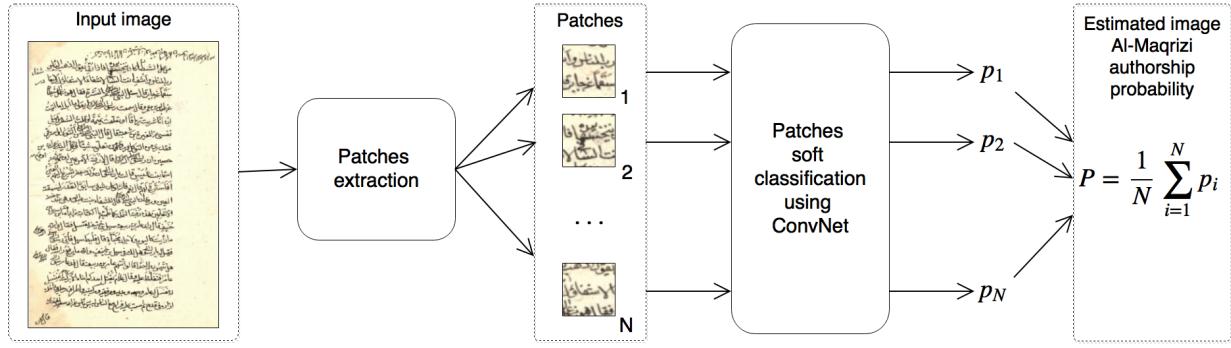


Fig. 1: al-Maqrizi authorship classification pipeline

- eight manuscripts not written by al-Maqrizi's hand (taken from the University of Michigan Hatcher Library Special Collections) are considered to be a negative example. These manuscripts were selected in such a way that the date and place of writing of each of them would be close to those of al-Maqrizi's *al-Khitat*, namely, the 14th and 15th centuries, Egypt and Syria.

To enhance robustness of the learning process the data obtained were divided into training and test 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.
- Test set: 1 al-Maqrizi's manuscript consisting of 14 pages and 3 non-al-Maqrizi's manuscripts each of which consists of 7 pages. Authors of these 3 manuscripts differ from authors of 5 negative examples in the training set.

It is important to note that we split our data set by the factor of the manuscript author. When the algorithm is learning on the training set, it can not see the authors in the test set. Thus, we enforce our method to infer general properties al-Maqrizi's handwriting not specific for particular manuscript. Number of non-al-Maqrizi's documents is chosen in such way that positive and negative classes in training and test sets are roughly balanced.

The goal of this study is to verify the authorship of the *al-Khitat* manuscript that consists of 32 pages.

III. METHOD

We consider author verification as a binary classification problem: the al-Maqrizi class designated as 1 and the non-al-Maqrizi class designated as 0. In this context our goal is to build a classification pipeline able to predict the probability (*soft classification*) that a 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) Soft classification of patches using convolution network.
- 4) Averaging predicted patches probabilities to produce overall candidate image of the probability of al-Maqrizi's authorship.

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

A. Image preprocessing

This step was done to bring all images to relatively same size and scale by performing the following steps:

- 1) Removing part of image within the text bounding box.
- 2) Resizing resulted image to resolution 700×500 .

These steps are reasonable enough since all images from our data set have approximately equal number of text lines and text bounding box aspect ratio.

Since this step is done only to unify our images data set we do not include it in the pipeline.

B. Patches extraction

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

1) *Sliding window based method*: This method splits an image into patches by a grid of fixed cells of the size 80×80 pixels with a stride 20 pixels. Figure 2 illustrates the idea.

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

- 1) Input image binarization using Otsu's filter [7].
- 2) Connected components extraction from a binarized image using the algorithm from [8].
- 3) Too small, too big and too stretched connected components filtering using several empirical rules, e.g.: major axis to minor axis ratio less than 10, minimum minor axis length greater than 3 pixels, etc.
- 4) Outlier connected components filtering using the DBSCAN clustering algorithm [9], [10].
- 5) Extracting remaining connected components bounding boxes from the source image and resizing them 28×28 pixels size.

Example of connected components based patches show on figure 3.

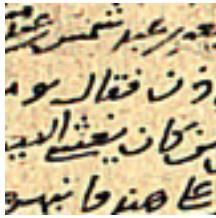


Fig. 2: Sliding window patch example



Fig. 3: Connected components patches example

As one can see, connected components-based patches usually consist of one or few letters thus providing a high robustness for different image scale and size in contrast to fixed-size sliding window patches (since all images been preprocessed this features is not important in our case). However, sliding-window patches contain much more information: several symbols from multiple lines — a very important feature for authorship verification.

C. Deep convolution network

Denote x_i as the i -th patch obtained from an image in the previous step and y_i as binary label. Then for predicting the probability

$$p_i = P(y_i|x_i)$$

that the current patch belongs to al-Maqrizi's hand we need to train a discriminative model. As a robust and powerful classification algorithm deep convolution networks [4], [5] are chosen.

In a case of sliding window patches we apply Alexnet type of network [6]. Architecture of this network described in table I.

TABLE I: Sliding window patch convolution network

type	patch size/stride	output number
convolution	11 / 4	96
local response norm		96
max pool	3 / 2	96
convolution	5 / 1	256
local response norm		256
max pool	3 / 2	256
convolution	3 / 1	384
convolution	3 / 1	384
convolution	3 / 1	256
max pool	3 / 2	256
fully connected		4096
dropout (50 %)		4096
fully connected		4096
dropout (50 %)		4096
fully connected		2
softmax		2

As an input network takes 80×80 RGB image. All activation functions in convolution and fully connected layers are rectified linear units (ReLU). This deep network was trained by stochastic gradient descent method with Nesterov momentum, initial learning rate is 0.01 and it decrease policy is step down.

With connected components patches we use GoogLeNet type of deep convolutional network [11]. Its architecture is provided in table II.

TABLE II: Connected components patch convolution network

type	patch size/stride	output number
convolution	7 / 2	64
max pool	3 / 2	64
local response norm		64
convolution	1 / 1	64
convolution	3 / 1	192
local response norm		192
max pool	3 / 2	192
inception		256
inception		480
max pool	3 / 2	480
inception		512
convolution	1 / 1	128
fully connected		1024
dropout (70 %)		1024
fully connected		2
softmax		2

As an input network takes 28×28 RGB image. All activation functions in convolution and fully connected layers are rectified linear units (ReLU). Inception layer is a combination of several convolution and pooling layers; for more details see [11]. Learning method for this convolutional network is stochastic gradient descent with Nesterov momentum and initial learning rate 0.01 with step down decrease policy.

IV. RESULTS AND DISCUSSION

We have experimented with both types of patches extraction and corresponding convolutional networks.

A. Deep convolutional network learning

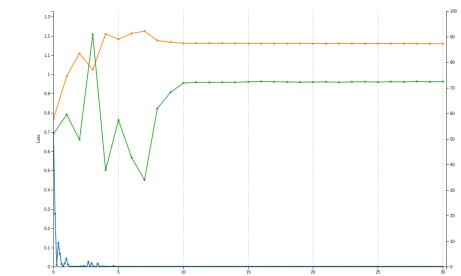


Fig. 4: Sliding window patch convolution network learning process.

Figure 4 demonstrates values of loss function on training (blue curve) and test (green curve) sets depending on the learning epoch for sliding window patch convolution network. Also one can see accuracy (orange curve) on the test set. Its final value is near 87%. For connected components patch convolution network accuracy is worse — 80%.

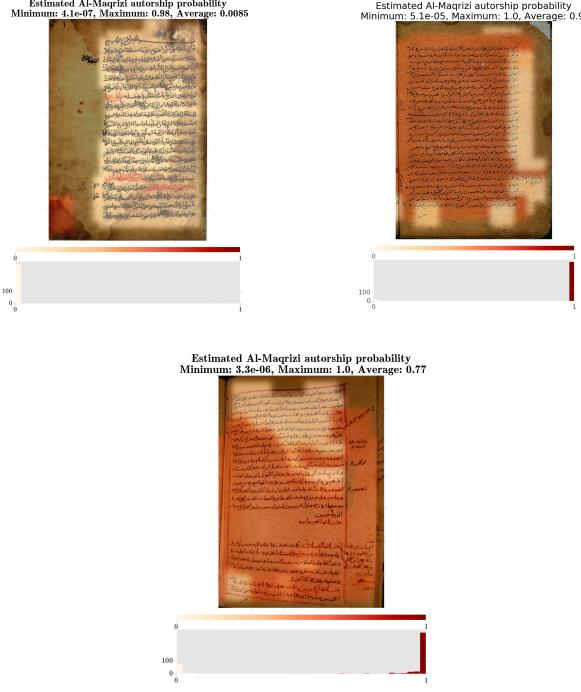


Fig. 5: Sliding window patches al-Maqrizi authorship classification example: non-al-Maqrizi manuscript page (left), al-Maqrizi manuscript page (right) the transcript page from *al-Khitat* (bottom). Patches probabilities are visualized by using white-red colors (corresponding to 0-1 classes) on the top of the original image. Additionally, a probabilities histogram is shown below.

B. Manuscript classification

To assess quality of the classification pipeline we have used only images from the test set, since they are the only ones not used in the learning process.

Figure 5 demonstrates the classification results for two images from the test set: one from al-Maqrizi class and one from non-al-Maqrizi class. As one can see, both of them can be classified quite confidently as having al-Maqrizi as the author with the probability 0.0085 and 0.94 respectively. Also this figure shows the classification result for an *al-Khitat* manuscript page. It gives 0.77 probability of al-Maqrizi being the author.

Regarding the entire test set classification for sliding window patches with decision threshold 0.5 we obtained precision 0.99 and recall 0.92. The method based on connected components patches is less robust: it generates many false positive predictions. This approach has a great potential, however, its improvement requires many more training examples. Thus convolutional network learned on connected components patches is a promising field for further research.

By using sliding window patches with deep network method we have arrived at the conclusion that the mean probability of al-Maqrizi's authorship of the entire set of the *al-Khitat* pages

equals 0.86.

V. CONCLUSION

The joint study of al-Maqrizi's "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. The results achieved thus far 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.

ACKNOWLEDGMENT

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.

REFERENCES

- [1] N. Gardiner and B. F., "A recently discovered holograph fair copy of al-maqrizi's *al-mawa'iz wa-al-i'tibar fi dhikr al-khitat wa-al-athar* (michigan islamic ms 605)," *Journal of Islamic Manuscripts*, vol. 2, pp. 123–131, 2011.
- [2] M. Bulacu, L. Schomaker, and A. Brink, "Text-independent writer identification and verification on offline arabic handwriting," *Proc. 9th International Conference on Document Analysis and Recognition, ICDAR*, pp. 769–773, 2007.
- [3] D. Fecker, A. Asi, W. Pantke, V. Mrgrner, J. El-Sana, and T. Fingscheidt, "Document writer analysis with rejection for historical arabic manuscripts," *Proc. 14th international Conference on Frontiers in Handwriting Recognition, ICFHR*, pp. 743–748, 2014.
- [4] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, no. 521, pp. 436–444, 2015.
- [5] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. of the IEEE*, pp. 2278–2324, 1998.
- [6] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.
- [7] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, pp. 285–296, 1975.
- [8] C. Fiorio and J. Gustedt, "Two linear time union-find strategies for image processing," *Theoretical Computer Science*, vol. 154, no. 2, pp. 165–181, 1996.
- [9] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *Proc. of Second International Conference on Knowledge Discovery and Data Mining*, vol. 96, no. 34, pp. 226–231, 1996.
- [10] O. Granichin, V. Volkovich, and D. Toledo-Kitai, *Randomized Algorithms in Automatic Control and Data Mining*. Heidelberg New York Dordrecht London: Springer-Verlag, 2015.
- [11] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.