Curation of Historical Arabic Handwritten Digit Datasets from Ottoman Population Registers: A Deep Transfer Learning Case Study

Yekta Said Can and M. Erdem Kabadayı
College of Social Sciences and Humanities
Koç University
İstanbul, TURKEY
ycan@ku.edu.tr, mkabadayi@ku.edu.tr

Abstract—With the increasing number of digitization efforts of historical manuscripts and archives, automatic information retrieval systems need to extract meaning fast and reliably. Historical archives bring more challenges for these systems when compared to modern manuscripts. More advanced algorithms, archive specific methods, preprocessing techniques are needed to retrieve information. Cutting-edge machine learning algorithms should also be applied to retrieve meaning from these documents. One of the most important research issues of historical document analysis is the lack of public datasets. Although there are plenty of public datasets for modern document analysis, the number of public annotated historical archives is limited. Researchers can test novel algorithms on these modern datasets and infer some results, but their performance is unknown without testing them on historical datasets. In this study, we created a historical Arabic handwritten digit dataset by combining manual annotation and automatic document analysis techniques. The dataset is open for researchers and contained more than 6000 digits. We then tested deep transfer learning algorithms and various machine learning techniques to recognize these digits and achieved promising results.

Index Terms—numeral spotting, historical document analysis, convolutional neural networks, deep transfer learning, handwritten digit recognition, dataset curation, page segmentation

I. INTRODUCTION

Historical archives contain valuable information about historical, social, and economic aspects of the past. These archives’ digitization processes gain pace in recent decades, including non-European handwritten archival collections [1]. Increased amounts of digitized archival information produce new challenges for researchers who work with historical archival data to apply novel automatic document analysis systems. The most commonly used techniques can be listed as layout analysis, keyword, symbol and number spotting, handwritten text recognition (HTR) and optical character recognition (OCR) [2].

Although there is plenty of research for analyzing modern documents and several open datasets, the number of historical document analysis studies is limited. The reason could be the challenges encountered when analyzing historical documents such as degrading documents, digitization errors, and different layout types. [3]. Different countries and cultures, e.g., Ottoman, Arabic, Urdu, Kurdish and Persian, use Arabic scripts in their manuscripts. However, a relatively small number of studies have been carried out to recognize Arabic text compared to the number of Latin, Chinese, and Japanese text recognition studies [4]. The possible reasons for this could be the cursive nature of Arabic scripts, dots and diacritics, different writing styles and the limited number of Arabic databases [3], [4].

There are plenty of studies for recognizing English handwritten digits, which reported high identification accuracies [5], [6]. Recently, several number spotting [7] and handwritten digit recognition schemes are proposed for Arabic scripts on different datasets ([8]–[10]). They achieved accuracies of more than 90%. However, these studies used datasets that are created lately, and they do not suffer from the historical documents’ mentioned problems.

We aim to detect individuals, households in populated places and recognize the age of individuals in the first series of Ottoman population registers. This is important because just by reading the assigned order of registered males in households, starting from the first until the last, we can make a substantial contribution to a hitherto unsolved debate in historical demography in Ottoman studies. The average number of people living in Ottoman households, and regional and ethno-confessional differences in household sizes are disputed subjects. It is commonly, yet without backed by any convincing empirical base, assumed that average Ottoman households had five members. On the other hand, it is a well-researched and overall accepted claim that polygamy was only marginally practiced. Therefore, when we can count the total number of males in households simply by multiplying by 2, we can estimate total members of households by far more reliably than any claim in the literature. By matching the order of the males with their ages, we can even further enrich our insights into the age
TABLE I
COMPARISON OF DATASETS FOR ARABIC HANDWRITTEN DIGIT RECOGNITION

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size (# of Images)</th>
<th>Origin Country</th>
<th>Historical</th>
</tr>
</thead>
<tbody>
<tr>
<td>HODA</td>
<td>70000</td>
<td>Iran</td>
<td>X</td>
</tr>
<tr>
<td>ADBase</td>
<td>70000</td>
<td>Egypt</td>
<td>X</td>
</tr>
<tr>
<td>CMATERDB 3.3.1</td>
<td>3000</td>
<td>India</td>
<td>X</td>
</tr>
<tr>
<td>Our Dataset</td>
<td>6000</td>
<td>Ottoman Empire 1840-1860</td>
<td>✓</td>
</tr>
</tbody>
</table>

structure of households and based upon the age distribution, we can project family formation types, such as whether they are multigenerational families or not. In this study, we first manually annotated the very first series of population registers of the Ottoman Empire conducted in the mid-nineteenth century. We then spotted and recognized the Arabic numbers and individuals and their pixelwise positions automatically. By automatically cropping the positions of individuals and numerals and merging them with manually read ground truths, we created the first historical Arabic handwritten open dataset. It includes more than 6000 handwritten digits and it is publicly available. We then tested commonly applied classifiers such as Random Forest, Multilayer Perceptron, Logistic Regression, Convolutional Neural Networks. Lastly, we applied deep transfer learning techniques from modern Arabic datasets to our historical dataset.

Fig. 1. Example of a figure caption.

The rest of the paper is organized as follows. The literature on modern Arabic handwritten digit datasets and Arabic digit recognition studies are provided in Section 2. The structure of the Ottoman population registers is presented in Section 3. We discussed how the dataset is created and used machine learning classifiers in Section 4. In Section 5, the experimental results and discussion are presented. We discuss the conclusion and future works of this study in Section 6.

II. RELATED WORKS

After spotting the numbers in the manuscripts, they must be recognized for retrieving information from the historical archives. There are several studies for recognizing Arabic digits in modern documents [10]. However, these studies used recently created datasets (see Table I) and can not be compared with research on historical documents. The HODA dataset [12] was created for developing Persian (an Arabic
handwritten digit recognition systems. It has 60000 training and 10000 test images, but some digits are different from classical Arabic digits. Another dataset is ADBase [13] created in Egypt. It has 60000 training and 10000 test images. The CMATERDB 3.3.1 dataset [14] was created at the Jadavpur University, India. It has 3000 images. All these datasets have been created in recent decades and they are tested with a variety of machine learning algorithms.

Different CNN architectures were tried with the HODA dataset. Lekhal et al. achieved 97.38% accuracy by using CNN + LeNet architecture [15]. Researchers achieved over 99% accuracies with CNN + CapsNet and CNN + AlexNet [16], [17]. CNN + LeNet architecture was also tested with ADBase and El Sawy et al. obtained 88% accuracy [10]. CNN and CNN + Boltzmann Machine classifiers were tested on CMATERDB 3.3.1 dataset and over 99% accuracies were obtained [18], [19]. However, to the best of our knowledge, there is not any publicly available historical Arabic handwritten digit or letter dataset. Our dataset will contribute to the literature on this aspect.

III. STRUCTURE OF THE POPULATION REGISTERS

We worked with the Nicaea district registers, NFS.d. 1452 and 1454 (see Figure 1 for a sample page of a register). They are publicly accessible at the Turkish Presidency State Archives of the Republic of Turkey—Department of Ottoman Archives in jpeg image format. We attempt to develop an automatic information retrieval technique for population registers from different empire precincts of the mid-nineteenth century. The registers comprise of extensive demographic information on the male population in the households, i.e., names, family relations, occupations and ages. The archives have around 11,000 population registers. The registers were indexed and gradually become available for research purposes since 2011. The height and width of the digitized page in the records were 2210 x 3000 pixels. There are different object types in these registers. The first one is the symbol that marks the beginning of a populated place (see Figure 1). It marks the end of a previous village and the start of a new one and we can use it to assign people to their villages or neighborhoods. Other object types are households or individuals counted in the census. These objects contain demographic information about individuals. If an individual is the first person of a household, there are two numerals (right and left) that show the id of the household and individual in the populated place at the top of the object. Otherwise, they put only one number on the left top of the object showing the number of the individual inside the populated place. At the bottom of all objects, the age of individuals appears (see Figure 1).

IV. METHODOLOGY

A. Manual Annotation of Ottoman Population Registers

This study is a preliminary work from an ongoing larger project. In our UrbanOccupationsOETR project, we reached an unprecedented scale of manual data entry and indirectly data annotation possibilities. In the first four years, the project
team located, copied, and read a total of 67 registers belonging to 96 locations, 9 of which were cities and around 20 of which were towns. The number of individuals in the populated places is also manually entered which could be used to match automatically detected individual pixel positions in the page images. Our database manager has set up and maintained a purpose-built relational database using and constantly updating data entry forms via MS Access. From those 67 registers, our team read and entered historical demographic information of in total of 167,902 individual males living in 69,242 households. We demonstrated the data entry process into our relational database for the city of Bursa (see Figure 2).

The entry of population registers for one city, allowed Erunal to come up with the very first attempt to estimate age structure and mortality rates of mid-nineteenth century Bursa. In our ongoing project, our manual entry would allow us to conduct similar studies on urban historical demography for around nine cities. However, by connecting our manually entered data, and therefore our annotation, with the thousands of images with automatic documentation analysis techniques, as we suggest here, we will have an unprecedented training set to further increase the accuracy of our deep transfer learning. In doing that, we will get closer to distant read thousands of population registers, which are digitally available at the Ottoman archives. Just to reiterate, the dataset of this study had around 6,000 digits. Our project has covered 167,000 individuals in 67 registers. The Ottoman archives have around 11,000 population registers just from the nineteenth century.

B. Automatic Page Segmentation and Numeral Spotting System

1) Automatic Page Segmentation and Individual Sorting System: As shown in Figure 1, individuals are written in clusters in the population registers. We first developed a page segmentation system for counting and sorting individuals (see our previous page segmentation study [20] for more details). Since the manual annotation sorts the individuals in the populated places, we automatically segmented individuals and sorted them to match coordinates of images with the manually annotated ground truth.

We first created a dataset for using the dhSegment toolbox [21]. We labeled individuals and populated place start objects with different colors. The start of a populated place object is colored with red, whereas the individual registers were marked with green. We marked 122 pages with the described labels. All of them belonged to the Nicaea district. An example of an original image and a labeled version are shown in Figure 3.

After creating the dataset, we trained a CNN model by using the dhSegment toolbox. It uses pretrained Resnet50 architecture. It employed L2 regularization with 106 weight decay, Adam optimizer [22] and Xavier initialization [23]. In order to overcome the lack of diversity problem, Batch renormalization [24] was used. It downsized images and divided them into 300 to 300 patches for fitting better into the memory and providing batch processing support. Different data augmentation methods such as rotation (from 0.2 to 0.2 rad), scaling (coefficient from 0.8 to 1.2), and mirroring were applied. For each class, the toolbox created a binarized matrix showing the probabilities that a pixel belonged to them. Using these matrices, we connected the pixels and created rectangles via the connected component analysis tool [21] (see Figure 4). We recorded the coordinates, object types of these rectangles in a csv file by writing new Python scripts.

We further sorted the individual objects inside populated places to match the manually annotated ground truths. We first created rows for each page. Rows are defined with the elements that have similar heights. We sorted rows by their
average heights. The elements inside rows are sorted by using their widths. Due to the properties of the Arabic language, the element on the left side is the first element. We sorted the individuals inside rows by abiding this rule. In this way, the individuals are assigned id numbers to match the ground truth.

2) Numeral Spotting System: In the studied registers, the numerals are written in red. By taking advantage of this property, we developed an automatic numeral spotting system (see our previous numeral spotting and recognition study [3] for more details). We then determined which individuals these numbers belong to. We lastly determined whether the number is a household number, individual number or age by examining its position in the individual rectangle and matched the manually annotated ground truth.

As mentioned before, the numerals in the studied registers are written in red. For spotting the numerals more straightforwardly, we applied a red color filter on the manuscripts. We transformed the JPEG pictures from RGB representation to HSV. We determined the upper and lower thresholds of the red color filter by trial and error. Lower HSV thresholds were determined as (170;70;50), whereas the upper HSV thresholds were selected as (180;255;255). An example original image of the register and the filtered one is shown in Figure 5. We chose the mask background color as black.

We created the second annotated dataset to train a CNN, which is a number dataset. It has two classes. The first one is the background, which is the area other than numbers and it is marked with black. The second one is the numeral region and we marked them with green. We marked Nicaea district registers with the described labels, which include approximately 5000 numbers (comprising one-digit, two-digit and three-digit numbers). The annotated numbers in the sample original image are presented in Figure 6. We used the dhSegment tool to train a numeral spotting CNN model. The pixels are connected and surrounding rectangles are created (see Figure 7). The coordinates of these rectangles are saved to a csv file.

All numbers are matched with detected individuals and their ids by using their coordinates in pages. Their position in an individual box determines the type of number (whether it is an individual id, household id or age). If it is at the bottom, it is age, if it is at the left top, it is an individual id, and if it is at the right top, it is a household id (see Figure 1). In this way, all numbers are matched with manually annotated ground truth.
C. Creating the Dataset

Using the page segmentation and numeral spotting systems, all numbers are matched with the manually annotated ground truth along with their coordinates in pages. We automatically cropped and named the cropped numbers as RegisterName_PageNumber_IndividualNumber_GroundTruth.jpeg. In order to convert the numbers into digits, the two-digit numbers are automatically divided into two. Recognized digits are automatically added to the datasets. Three-digit numbers are discarded. The dataset was controlled manually and cleaned. It has more than 6000 digits. Five thousand images are divided into the training folder, and the remaining 1000 images are divided into the test folder. The number of images of different digits is not balanced. Researchers should report F_measure along with accuracy or apply techniques for dealing with class imbalance such as randomly undersampling and SMOTE. The dataset can be accessed at https://urbanoccupations.ku.edu.tr/public-datasets/. We created three different versions, namely greyscale 28x28 and 64x64 versions and colored 28x28 version. The resizing of all images to the mentioned dimensions and greyscale conversions are carried out by using Python scripts.

D. Digit Recognition System

1) Locally Trained Digit Recognition Models: We first tested locally trained and tested classifiers on this new dataset. The applied traditional classifiers were Gradient Boost, Random Forest and Multilayer Perceptron. The default parameters of the Python Keras library were used. CNN algorithm is used as a deep learning algorithm and the CNN architecture was created and tested on this dataset (see Figure 8).

2) Pretrained Digit Recognition Models with HODA and ADBase Datasets: We also investigated the effect of pretrained networks on our digit recognition performance. The HODA dataset and ADBase datasets comprise of 70000 images. We trained a CNN model by using these training sets. After that, we save the model parameters and use them as starting parameters of the CNN model while training our dataset. We called the CNN models trained with the HODA dataset as HODANet and with the ADBase dataset as ADBaseNet. We selected these modern Arabic handwritten digits because of their size and similarity to our task. Although they are not historical datasets, our system could take advantage of model parameters learned from these datasets. We expect this technique will fasten the training and improve the performance of our digit recognition system.

3) Digit Recognition using Deep Transfer Learning from HODA and ADBase Datasets: We further applied deep transfer learning from the HODA and ADBase datasets. The output layer of the trained models with HODA and ADBase datasets is removed in this technique. We used the last layer before the output layer for features. It has (128,1) dimensions. The trained CNN models (with HODA and ADBase datasets) took our test set of the dataset as input and output these 128 features. These extracted features (the output) were then fed into machine learning classifiers to recognize digits. K Nearest Neighbours, Logistic Regression, Multilayer Perceptron, Random Forest and Support Vector Machine classifiers were selected from literature for this purpose.

V. EXPERIMENTAL RESULTS AND DISCUSSION

A. Digit Recognition Results using Locally Trained Models

We trained and tested selected machine learning algorithms in our dataset. The digit recognition accuracies are presented in Table II. The CNN architecture outperforms the other classifiers. It is expected because CNN is a deep learning classifier that is suitable for image processing applications. Random Forest was the most successful classifier among traditional algorithms.

B. Digit Recognition Results using Pretrained Models

The results obtained by using pretrained HODA and ADBase models are presented in Table III. We can see that although using pretrained models did not increase our digit recognition performance, it decreased the number of epochs needed to converge (see Figure 9). It reaches the maximum accuracies quickly, which shows that it decreases training...
TABLE III

THE DIGIT RECOGNITION ACCURACIES OBTAINED WITH CNN AND PRETRAINED MODELS ARE USED FROM KNOWN MODERN DATASETS.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Model Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN + HodaNet</td>
<td>99.60</td>
<td>0.009</td>
</tr>
<tr>
<td>CNN + ADBaseNet</td>
<td>98.70</td>
<td>0.0389</td>
</tr>
</tbody>
</table>

C. Digit Recognition Results using Deep Transfer Learning from HODA and ADBase Datasets

We obtained results by removing the output layer of the CNN architecture and using the previous layer parameters for feature extraction in our test set (128 features). When we used different classifiers for these extracted features, we achieved lower performance than the locally trained CNN model. We presented accuracy, f_measure and area under the ROC curve results in Tables IV and V. Multilayer Perceptron (MLP) achieved the best results when Deep Transfer Learning (DTL) is applied from both datasets. The reason for the lower performance is the difference of features in historical and modern datasets which proves the necessity of historical datasets for studying archival data. Using model parameters of the HODA dataset achieves lower performance with all classifiers because some digits are different from the classical Arabic scripts in Persian Arabic scripts (see [3]).

VI. CONCLUSION

In this study, the first series of population registers are manually annotated. We then applied page segmentation and number spotting algorithms and merged the manually annotated ground truth with pixelwise positions of the numbers. We created the first historical Arabic handwritten datasets and made it publicly accessible. We believe that this dataset will help researchers working with historical Arabic handwritten
results. We further applied different types of machine learning classifiers such as Multilayer Perceptron, CNN and Random Forest. We also tried Deep Transfer Learning techniques from big Arabic handwritten digit datasets, namely HODA and ADBase. We achieved the best results when training and testing CNN with our dataset. Transfer learning methods also achieved over 90% accuracy, but the reason for slightly lower performance is the difference of characteristics in historical and modern datasets, which proves the necessity of historical datasets for studying archival data. As future works, we plan to increase the number of instances in this dataset by automatically analyzing more population registers and add an Arabic handwritten occupation dataset.

REFERENCES


