electronics MDPI

Article

Automatic Estimation of Age Distributions from the First Ottoman Empire Population Register Series by Using Deep Learning

Yekta Said Can * and M. Erdem Kabadayı

College of Social Sciences and Humanities, Koç University, Rumelifeneri Yolu, Sarıyer, Istanbul 34450, Turkey; mkabadayi@ku.edu.tr
* Correspondence: ycan@ku.edu.tr

Abstract: Recently, an increasing number of studies have applied deep learning algorithms for extracting information from handwritten historical documents. In order to accomplish that, documents must be divided into smaller parts. Page and line segmentation are vital stages in the Handwritten Text Recognition systems; it directly affects the character segmentation stage, which in turn determines the recognition success. In this study, we first applied deep learning-based layout analysis techniques to detect individuals in the first Ottoman population register series collected between the 1840s and the 1860s. Then, we employed horizontal projection profile-based line segmentation to the demographic information of these detected individuals in these registers. We further trained a CNN model to recognize automatically detected ages of individuals and estimated age distributions of people from these historical documents. Extracting age information from these historical registers is significant because it has enormous potential to revolutionize historical demography of around 20 successor states of the Ottoman Empire or countries of today. We achieved approximately 60% digit accuracy for recognizing the numbers in these registers and estimated the age distribution with Root Mean Square Error 23.61.

Keywords: line segmentation; convolutional neural networks; page segmentation; Arabic document processing; projection profiles; digit detection and recognition

check for updates

Citation: Can, Y.S.; Kabadayı, M.E. Automatic Estimation of Age Distributions from the First Ottoman Empire Population Register Series by Using Deep Learning. *Electronics* **2021**, *10*, 2253. https://doi.org/10.3390/electronics10182253

Academic Editor: Shyam Prasad Adhikari

Received: 20 August 2021
Accepted: 12 September 2021
Published: 13 September 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

We have been living in written cultures for ages, and we not only produce vast amounts of documentation, but we also are governed and ruled by them. The necessity of analyzing documentation and understanding and processing their content are perpetual. In the past, processing the information and correspondence kept in manuscripts was performed manually because of the lack of comprehensive and high-quality digitized datasets where an automatic method could be employed. Because of the rarity of high-quality digital scanning solutions and devices with high-storage capability, transforming and saving manuscript images from paper form to digital form was difficult. Recently, this job has become more evident due to dramatic progress in digital scanning and storage solutions.

Nowadays, there are many digitized historical documents in Arabic script in the national libraries and archives around the world, thanks to the above-mentioned advances in technology. Investigating data and retrieving information manually are costly and challenging. Therefore, an automatic method is needed to process these documents rapidly. Processing historical manuscripts is an up-to-date research topic that has seen dramatic growth recently [1–3]. However, historical Arabic document processing is a difficult research issue. The reasons could be listed as the complex nature of Arabic script compared to other scripts, and fragility of ancient documents, which are subject to degradation [4].

When recognizing handwritten documents, segmenting the document images into their primary objects, words, and text lines is a highly complex research issue in Arabic manuscripts because of the various problems faced in both word and text line segmentation processes. The main difficulties in the segmentation process of Arabic lines are overlapping words, very close neighboring text lines, and over the same text line or on the page between lines, the variances of the angle skew [5]. In this study, we worked with the first population registers of the Ottoman Empire that were conducted in the 1840s. The coverage of these registers is the entire Ottoman Empire in the mid-nineteenth century, which comprises the areas of around twenty successor countries of modern Southeast Europe and the Middle East; they are written in Arabic scripts. In these censuses, clerks created manuscripts without using hand-drawn or printed tables and there was not any pre-determined page structure. In the last line of demographic data of individuals, the ages are written and we aim to recognize them.

Retrieving age information from historical manuscripts has huge potential for the field of historical demography. Although the Ottoman population registers convey minute and diverse detail on individuals such as names, family relations, occupations, physical appearance, and body height for the purposes of historical demography, age data and its distribution are the most telling and the most rewarding ones. The age structure of any population in any given time is a crucial marker of demographic transition [6,7]. Furthermore, age heaping, a directly related phenomenon to registered age distribution, has vast potential for historical demography studies with its connection to human capital formation [8]. A computerized age information retrieval method for handwritten Arabic numbers tailored for these Ottoman population registers does not only have massive potential to revolutionize historical demography of around 20 countries of today, but it also has the potential to improve HTR methods for historical and modern handwritten documentation in Arabic.

In this study, we developed software that automatically segments pages and detects lines in these objects from the population registered in Ottoman populated places or settlements. We automatically obtained the last line, which includes the age information and estimated the age distribution in these historical registers by recognizing the numbers. For this study, we focused on one location: Manisa town in western Anatolia in Turkey. We first employed a CNN-based page segmentation technique to retrieve demographic data of individuals by using the models developed in our previous studies [9,10]. After that, we used horizontal projection profile-based line segmentation to the demographic information of these detected individuals in these registers and obtained the age data in the last line. We further detected and recognized the ages and estimated the age distribution.

The structure of the paper is designed as follows. In Section 2, the related work for line segmentation and Arabic digit recognition studies will be examined. We describe the structure of the population registers in Section 3. Our methods for page segmentation, line detection and digit detection and recognition are described in Section 4. Experimental results and a discussion are presented in Section 5. We present the conclusion and future works of the study in Section 6.

2. Related Works

Digitization of historical archives and application of information retrieval methods on them have gained pace in recent decades, including non-European handwritten archival collections [11]. Some historical documents might have a tabular structure, which makes it easier to analyze the layout. Zhang et al. [12] developed a system for analyzing Japanese Personnel Record 1956 (PR1956) documents, which includes company information in a tabular structure. They segmented the document by using the tables and applied Japanese OCR techniques to segmented images. Kusetogullari et al. [13] proposed a new automatic historical handwritten digit detection and recognition framework named DIGITNET. To train this network, they created and used a new historical handwritten digit dataset (DIDA), which contains 250,000 single digits, 25,000 images with bounding box annotations, and 200,000 digit strings from the tabular Swedish historical handwritten documents of the 19th century. However, in most of the cases (as in our case) these archival documents do not have a tabular structure. Cheddad et al. [14] created a semi-annotated dataset from the Swedish Historical Birth Records from 1800 to 1840 and made it public for researchers working in this area. They evaluated some deep learning models for word spotting.

In the Arabic manuscript processing literature, a wide variety of segmentation techniques were reported. In [15], the authors started by removing outlier elements by using a threshold; then, the letters linked to two lines at the half distance are recognized and horizontally segmented. For detecting the lines, a rectangular neighborhood on a current element is centered and rises to contain particular conditions. The distance filtered elements to the corresponding lines are then allocated.

In another study, the authors developed an algorithm for detecting lines from Arabic handwritten documents by mentioning the problems of multi-touching and overlapping characters [16]. The unsupervised method depends on the analysis of a block covering. The authors first analyzed a statistical block that computes the specific number into vertical strips of manuscript decomposition. Then, they employed the fuzzy C means method, which accomplishes fuzzy-based line detection. Last, they assigned the blocks to their corresponding lines. They achieved 95% accuracy for detecting lines in Arabic manuscripts in their dataset.

In [17], the researchers applied morphological dilatation and projection profile techniques. In order to estimate the skew of the line, they used horizontal projection profiles. In every zone for smearing, they employed the slope, using dilation with adaptive structuring element to do the changes according to the zone, the slope, and the size. The big blobs are identified in the second stage with a recursive function that searches the cut point.

In [18], the projection profile method is again employed after joining cut characters and eliminating small elements to define the point of division within the horizontal projection profile; the curve of Fourier fitting is employed. The contour is employed for segmenting the baseline of the connected component, which permits defining the cut point between different neighbor lines. The nearest line is approximated by the curve of a polynomial that fits in the baselines of the pixels. As it could be seen from the literature, several techniques were applied to segment lines in Arabic manuscripts. Due to its convenience and wide usage, we selected a projection profile-based technique for our problem.

After detecting the numbers in the last line, they must be recognized for retrieving data from the population registers. Several studies have recognized Arabic digits in contemporary documents [19]. However, they employed recently created datasets that are clean and not complicated when compared to historical documents. The HODA dataset [20] was formed for creating Persian (an Arabic script-based language) handwritten digit recognition systems. It has 60,000 training and 10,000 test images, but some digits are different from classical Arabic digits. A similar and larger Farsi digit dataset was created in another study [21]. It has around 100,000 digits obtained from university and college students. Another dataset is ADBase [22], which was created in Egypt. It has 60,000 training and 10,000 test images. The CMATERDB 3.3.1 dataset [23] was created at 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 [24–26] and researchers achieved accuracies over 95%. CNN and CNN + Boltzmann Machine classifiers experimented on CMATERDB 3.3.1 dataset and over 99% accuracies were obtained [27,28]. However, to the best of our knowledge, there is not any publicly available historical Arabic handwritten digit or letter dataset. Training a deep learning model with these modern datasets and testing with historical digits did not yield high accuracies in the literature [29]. Therefore, we created a dataset containing over 6000 digits, which contributes to the literature on this aspect [9]. The dataset can be accessed at https://urbanoccupations.ku.edu.tr/historical-arabic-handwritten-digit-dataset/ (accessed on 13 September 2021). We then trained a CNN model and recognized the digits in this case study.

3. Dataset Description

The mid-nineteenth century Ottoman population registers resulted from an unprecedented governmental procedure, which aimed to record every male subject of the empire, irrespective of age, ethnic or religious affiliation, or military or financial status. They intended to have universal coverage for the male population. Government officials created these manuscripts without using hand-drawn or printed tables. Therefore, a predetermined page layout did not exist. Page structure can change in different districts, and structural variations occurred depending on the clerk's preferences in different registers. This research focused on the city of Manisa registers, with code names NFS.d. 2865, 2866, 2867, available at the Turkish Presidency State Archives of the Republic of Turkey, Department of Ottoman Archives, in jpeg format, upon request. We aimed to implement a method for recognizing text of similar registers from different regions of the Ottoman Empire conducted between the 1840s and the 1860s.

As we mentioned, these registers contain comprehensive demographic data on male members of the households, i.e., names, family relations, ages, and occupations. Females in the households were not recorded. The registers are provided for research purposes at the Ottoman State Archives in Turkey, as recently as 2011. Their total number is around 11,000. Until now, they have not been subject to any systematic study. Only individual registers were investigated in a piecemeal fashion. The size of the digital images of registers is 2210 × 3000 pixels. A sample register page is demonstrated in Figure 1.

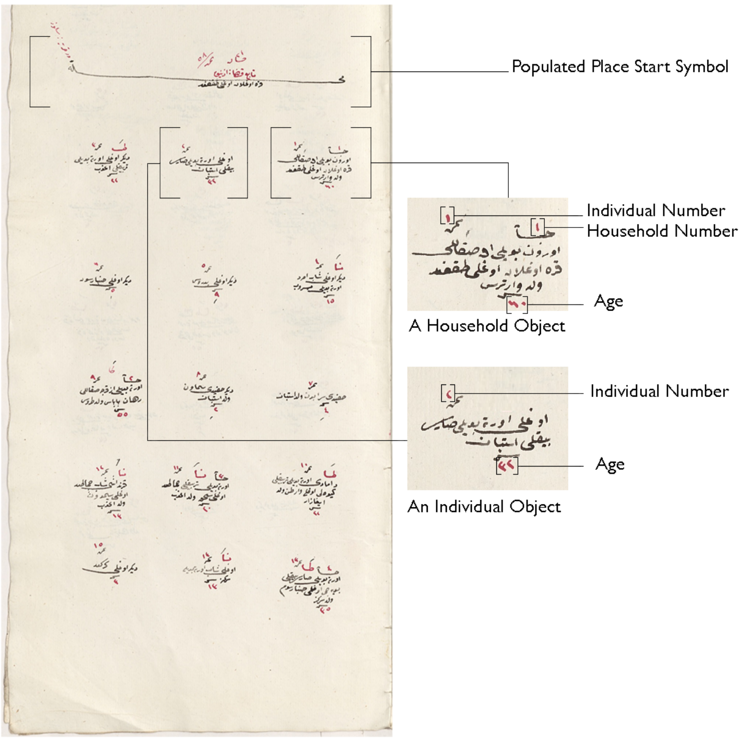

Figure 1. A sample register page and individual object are demonstrated. The individual and household numbers are written in the top of the individual data and ages are written in the last lines.

For estimating the age distribution, we selected three neighborhoods from Manisa city with respective numbers of registered persons: Sultan (120), Çarşı (68), and Gürhane (225). They contain a total of 413 people and their ages are estimated.

4. Automatic Age Estimation System

4.1. CNN-Based Object Detection Method

First, we developed a deep learning algorithm for detecting individuals in the population registers in our previous studies. We created a manually labeled dataset by using several registers and trained CNN models by using the dhSegment tool [30]. In the CNN-based dhSegment toolbox, paths use pretrained weights from well-known architectures such as Unet and Resnet50, where the system learns high-level features. They improve robustness and generalization. With the pretrained weights in the network, the training time and the number of parameters in the CNN architecture were reduced considerably [30].

We trained different models for different types of layouts. The first model was trained for registers with tightly placed individuals. The second model was trained for registers with loosely placed individuals. We used the former model for Manisa registers of this study. After we detected the individual objects in these registers, by using the pixelwise locations, we cropped the demographic data of individuals to be used for line detection algorithms. The detected individual objects can be seen in Figure 2. For more detailed information, our previous paper on object detection for these population registers could be visited [10].

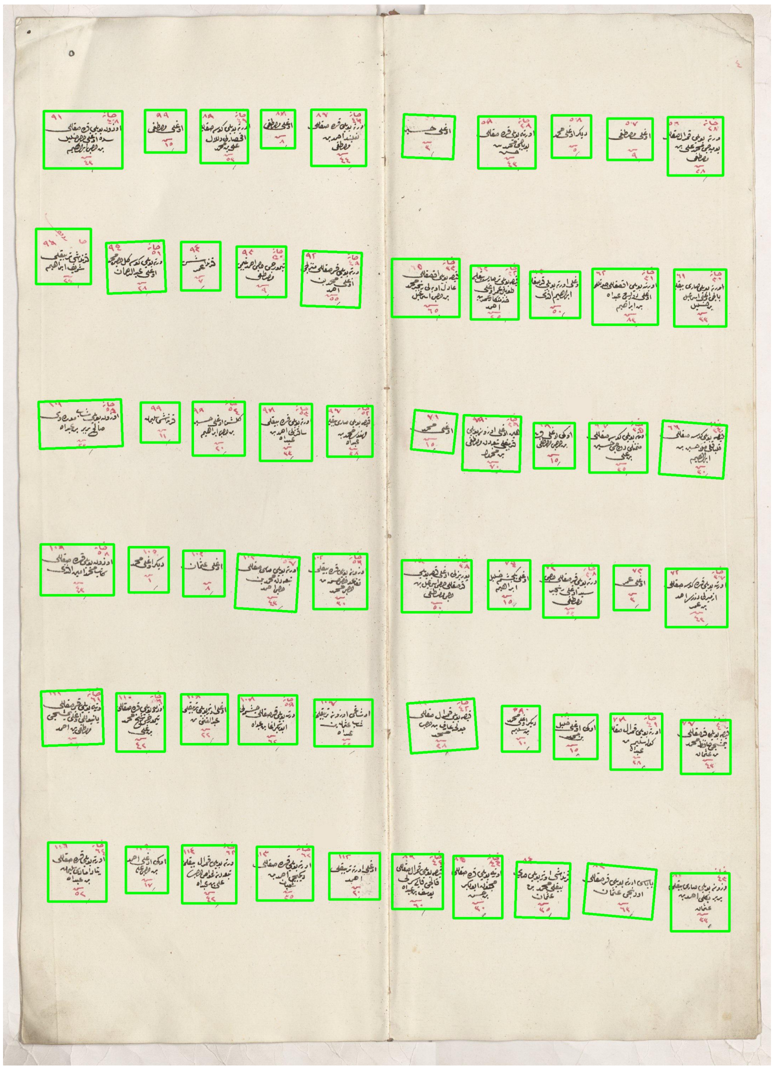

Figure 2. Detected individuals are bounded with green boxes. They include demographic data of individuals. As the pixelwise positions are detected, they are cropped for providing input to the line detection method.

4.2. HPP-Based Line Segmentation Method

One of the widely used methods for finding the line height of a document is by examining its horizontal projection profile. Horizontal projection profile (HPP) is the array of sum or rows of a two-dimensional image. Wherever there are more void areas between lines, we can observe more peaks. In order to give an idea of where should the segmentation between two lines can be employed, a sample HPP is provided in Figure 3.

In the first Ottoman population registers series, the ages are written in the last line of each entry. Therefore, we applied a peak detection algorithm for detecting the last peak to separate the age information in the last line. The Peakutils library of the Python programming language is used to find peaks. The void regions between lines appear as valleys, so we reverse the function and find peaks that correspond to valleys in the original HPP. We adapted the parameters of the Peakutils library to find the largest peak which divides the last line from the individual object. The peak threshold is selected as one-fourth of the maximum HPP and the minimum distance to look for a peak after a detected peak is selected as 10 datapoints by using trial and error. As we are searching for the last line, we crop everything under the last and largest peak in the reverse HPP, which provides us the age of the individual.

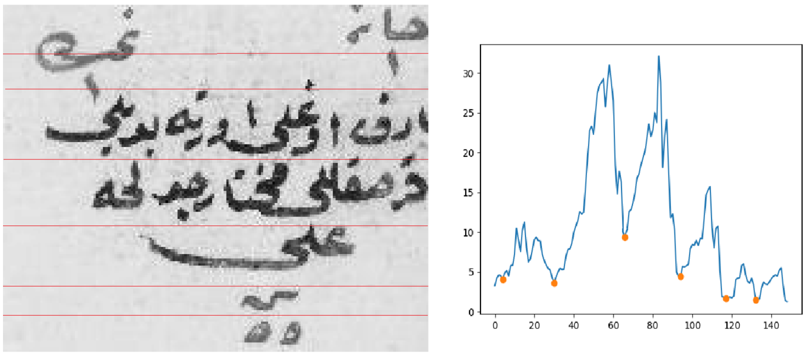

Figure 3. A sample segmented individual in the left and corresponding valleys in HPP in the right are shown.

4.3. Digit Detection and CNN-Based Digit Recognition System

After line segmentation, we automatically crop the last line, which contains the age of individual (see Figure 4). We detected the digits, recognized them and directly assigned the digit value if there is only one digit. If there are two digits, we combined the digits into two-digit numbers by making necessary calculations.

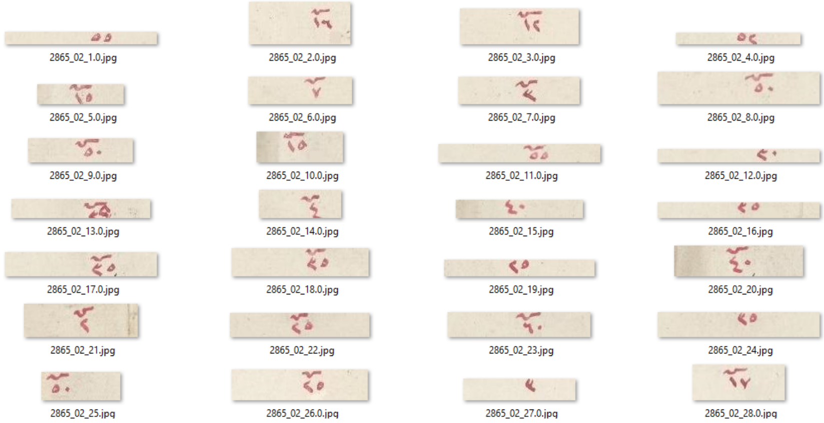

Figure 4. A subset of automatically cropped ages are demonstrated.

4.3.1. Digit Detection and Training Machine Learning Models for Digit Recognition

For detecting the digits, the OpenCV-Python library is used. It first found contours. From these contours, it created bounding rectangles. We first applied a basic filter that if it was too small, it was detected as noise and we eliminated these rectangles. The other rectangles were provided as input to the digit recognition system.

By using our publicly available historical Arabic handwritten digit dataset, we trained machine learning models. We chose four different algorithms: logistic regression (LR), a shallow Neural Network (one input and one output dense layer), a Deep Multilayer Perceptron (one output, one input and two hidden dense layers), and CNN. The first three algorithms are applied to the raw image by converting 2D matrices directly to 1D arrays. CNN is applied to the 2D images directly. These algorithms are selected because each one represents a different type. We used the same CNN architecture with our previous study [29] (see Figure 5). The accuracy and loss of training and test data with respect to epochs are provided in Figures 6 and 7. The models are saved as h5 file and the bounding rectangles are provided as input to these models and they give the predicted digits with their probabilities.

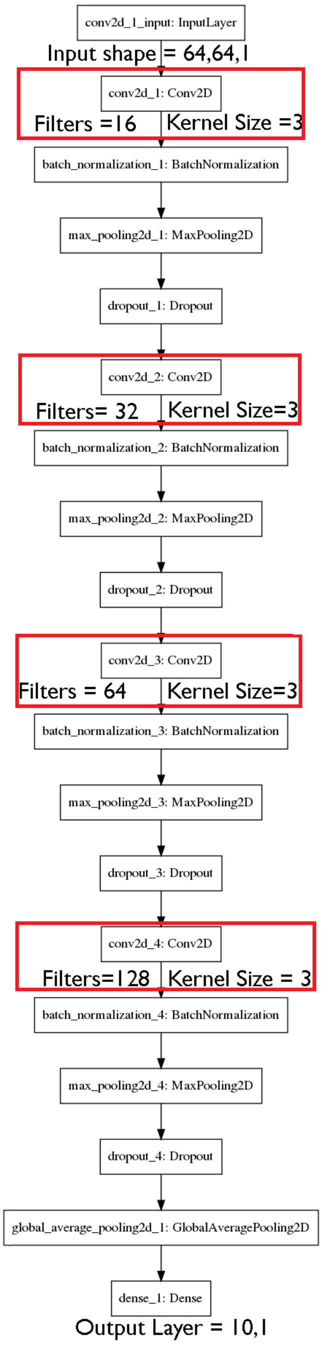
conv2d_1_input: InputLayer
Input shape = 64,64,1
conv2d_1: Conv2D
Filters =16
Kernel Size =3
batch_normalization_1: BatchNormalization
max_pooling2d_1: MaxPooling2D
dropout_1: Dropout
conv2d_2: Conv2D
Filters= 32
Kernel Size=3
batch_normalization_2: BatchNormalization
max_pooling2d_2: MaxPooling2D
dropout_2: Dropout
conv2d_3: Conv2D
Filters = 64
Kernel Size=3
batch_normalization_3: BatchNormalization
max_pooling2d_3: MaxPooling2D
dropout_3: Dropout
conv2d_4: Conv2D
Filters=128
Kernel Size = 3
batch_normalization_4: BatchNormalization
max_pooling2d_4: MaxPooling2D
dropout_4: Dropout
global_average_pooling2d_1: GlobalAveragePooling2D
dense_1: Dense
Output Layer = 10,1

Figure 5. The CNN architecture used for digit recognition.

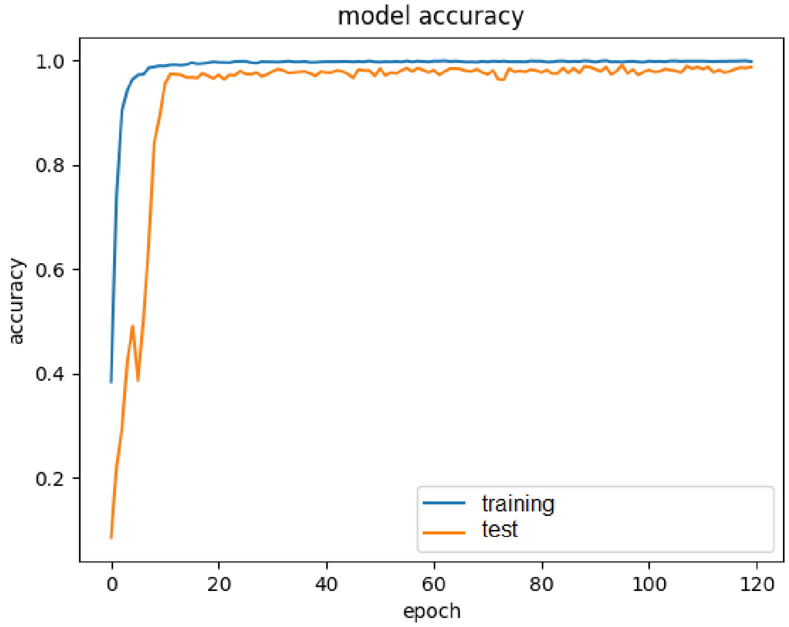

Figure 6. Training and test accuracies with respect to epochs.

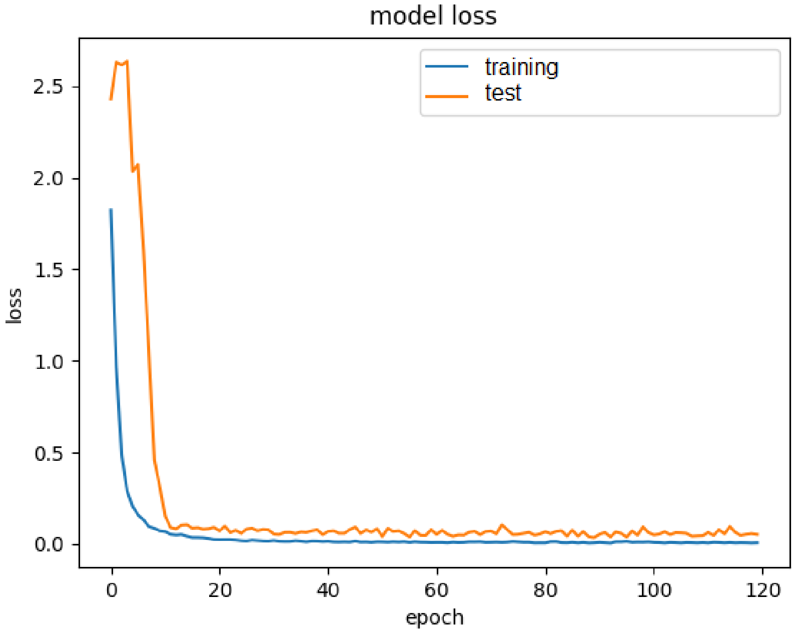

Figure 7. Training and test loss values with respect to epochs.

4.3.2. Transforming Digits into Numbers

The ages in this case study ranged between 0 and 83. Therefore, they are either one-digit or two-digit numbers. If we detect only one digit, its value is assigned as age. However, if there are two digits, we check their locations and we multiply the digit at the left side with ten and add the digit at the right side and assign this sum as the age of an individual. The predicted ages of individuals are computed in this way. We retrieved the ground truth age of individuals from our UrbanOccupationsOETR databases, which are manually entered by our project members, who are experts in Ottoman paleography. The metrics were calculated for each age entry and then we draw the histogram of ages and compared it with ground truth distribution.

5. Experimental Results and Discussion

In this section, we first provided our individual object detection results for Manisa population registers. After that, we presented last line detection accuracies for these registers.

5.1. Individual Object Detection Results

We tested the performance of our system on Manisa registers (more than 7000 individuals). We employed two different pretrained architectures, namely, Unet and Resnet50, and presented results by using both of them. In order to count individuals in the registers, we defined a high-level metric that can be calculated by dividing the predicted count errors over the ground truth count. We named this metric as Individual Counting Error (ICE).

$$ICE = || \frac{PredictedIndividualCount - Ground - TruthIndividualCount}{Ground - TruthIndividualCount} || \quad (1)$$

We used a tightly placed model for Manisa as its structure is more suitable for these models. The results are presented in Table 1.

Table 1. The individual counting error results are presented.

Pretrained Network	ICE (%)
Unet	0.3836
Resnet50	0.6536

As can be seen from Table 1, the counting errors are around 0.5%. Using a pretrained Unet architecture reduces the error percentages from 0.65% to 0.38%.

5.2. Line Detection Results

We evaluated the performance of our line detection system. We again used a high-level metric, which is the line detection accuracy. It could be calculated by dividing the number of correctly detected lines by the total number of lines. In our previous study, we detected all lines with detection accuracies of 80.30% for Manisa registers [31]. As we used only the last line in this study, when we examine the last line detection performance, we successfully detected the last lines with 100% accuracy. As the last line is more distant from the main body than classical line breaks, it makes sense that it could be detected with higher accuracies.

5.3. Digit Recognition Results

After the last lines are automatically cropped and the digits in the ages are detected via the OpenCV library, we provided these digits to the trained machine learning models and predicted the results. We compared the machine learning algorithms in two different ways: The first one is the digit recognition performance in the test set of our public dataset. We presented the results in Table 2. CNN outperforms the other algorithms in this comparison. The second way is to compare the performances of the trained and saved models for recognizing digits in the case study. We measure this performance by digit recognition accuracy and presented the results in Table 2. We further calculated digit detection accuracy, the root mean square error, average error metrics for best performing CNN architecture and presented them in Table 3. In our case study for digit recognition, we selected three neighborhoods from Manisa city, which contain 413 people and 739 digits in their ages. We calculated the digit detection accuracy by dividing the number of correctly predicted digits by the number of all digits. We predicted 60% of the digits correctly with the best machine learning model, which is CNN. When we examine the errors, 28% of the errors are caused by 0. In handwritten Ottoman, 0 is written with a dot and it could be easily confused with a noise in the document (see Figure 8). Therefore, it is also hard to detect. Another common error is the distinction between 2 and 3. They are quite similar in handwritten Ottoman (see Figure 8) and could be easily confused by our model.

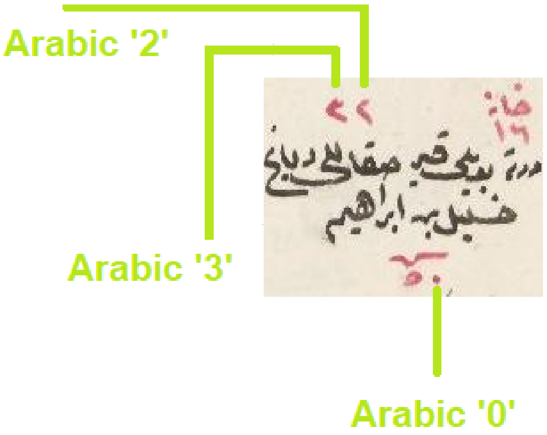
Arabic '2'
Arabic '3'
Arabic '0'

Figure 8. Arabic numbers 0, 2, and 3 from the population registers are shown. It could be seen that 2 and 3 are very similar and 0 is hard to detect.

Table 2. The comparison of machine learning algorithms in both recognition performance in the test set of our public dataset and the performances of the trained and saved models for correctly recognizing digits in the case study.

Algorithm	Digit Recognition Accuracy from the Test Set of the Public Dataset (%)	Correctly Recognized Digits in the Case Study (%)
Logistic Regression	42.21	10.33
Shallow Neural Network	32.76	21.36
Deep Multilayer Perceptron	32.28	20.27
CNN	98.35	58.18

Table 3. The digit detection results are presented.

	Digit Accuracy (%)	RMSE	Average Absolute Error
Unet	58.186	23.606	14.703

5.4. Age Distribution Estimation

The ultimate aim of this study is to estimate the age distribution of the selected three neighborhoods in the city of Manisa. After combining the predicted digits into numbers, we plotted the distribution of these numbers as a histogram (see Figure 9). The histograms are quite similar, but we can see that the number of people whose ages are between 0 and 20 is higher in the predicted numbers than the ground truth. The reason is that the difficulty in predicting Arabic '0' (which is a dot). The algorithm generally misses to detect if there is a digit in the case of zeros in the two-digit numbers and predicted them as one digit, which increases the 0–20 ages. Beside the zero problems, the algorithm successfully estimates the age distribution from the historical population registers.

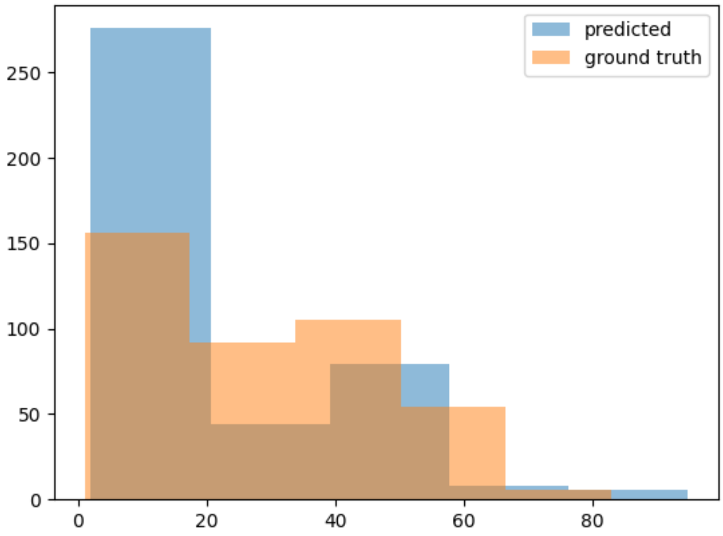

Figure 9. The age distributions of ground truth ages and predictions are demonstrated.

6. Conclusions

In this study, we first used a CNN-based layout analysis technique to detect individual objects in the first Ottoman population register series. Then, we employed a horizontal projection profile algorithm-based line segmentation to the demographic information of these detected individual objects for detecting the last line, which contains the age data. We further detected and recognized digits and converted them to numbers. We focused on Manisa register individual entries. We detected objects that include demographic data of individuals with approximately 0.5% error by using the CNN-based segmentation algorithm. We detected the last lines of information belonging to 413 individuals of our case study from Manisa with 100% success rate. We detected digits by using contour lines and finding bounding rectangles and recognized them with a CNN model trained with our public dataset. We achieved approximately 60% digit detection accuracy. We estimated the age distribution of the selected neighborhoods with promising accuracy. We plan to add word and letter detection systems as future works and develop a recognition system for these registers. This will reveal important demographic information from a wide geographical area in the nineteenth century and will have a significant interdisciplinary impact on historical demography.

Author Contributions: Y.S.C. is the main writer of the manuscript. He performed the curation and development of the dataset and of the software and conducted the analysis. M.E.K. organized the preparation of the archival sources and initial data gathering. He has provided historical context and information regarding late Ottoman population registers, and contributed to the conceptualization of the case study. Both authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the European Research Council (ERC) project: “Industrialisation and Urban Growth from the mid-nineteenth century Ottoman Empire to Contemporary Turkey in a Comparative Perspective, 1850–2000” under the European Union’s Horizon 2020 research and innovation program Grant Agreement No. 679097, acronym UrbanOccupationsOETR. M. Erdem Kabadayı is the principal investigator of UrbanOccupationsOETR.

Data Availability Statement: UrbanOccupationsOETR_hdr_Nicaea_6k is the first historical Arabic handwritten digit dataset. It is curated from the first series of Ottoman population registers conducted in the mid-nineteenth century. The dataset was controlled manually and cleaned. It has more than 6000 digits. 5000 are divided into the training folder, and the remaining 1000 images are divided into the test folder. The dataset can be accessed at https://urbanoccupations.ku.edu.tr/historical-arabic-handwritten-digit-dataset/ (accessed on 13 September 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kahle, P.; Colutto, S.; Hackl, G.; Mühlberger, G. Transkribus—A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; Volume 4, pp. 19–24.
2. Colutto, S.; Kahle, P.; Guenter, H.; Muehlberger, G. Transkribus. A Platform for Automated Text Recognition and Searching of Historical Documents. In Proceedings of the 2019 15th International Conference on eScience (eScience), San Diego, CA, USA, 24–27 September 2019; pp. 463–466.
3. Muehlberger, G.; Seaward, L.; Terras, M.; Oliveira, S.A.; Bosch, V.; Bryan, M.; Colutto, S.; Déjean, H.; Diem, M.; Fiel, S.; et al. Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *J. Doc.* **2019**, *75*, 954–976 . [CrossRef]
4. Ibn Khedher, M.; Jmila, H.; El-Yacoubi, M.A. Automatic processing of Historical Arabic Documents: A comprehensive Survey. *Pattern Recognit.* **2020**, *100*, 107–144. [CrossRef]
5. Ali, A.A.A.; Suresha, M. Efficient algorithms for text lines and words segmentation for recognition of Arabic handwritten script. In *Emerging Research in Computing, Information, Communication and Applications*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 387–401.
6. Lee, R. The Demographic Transition: Three Centuries of Fundamental Change. *J. Econ. Perspect.* **2003**, *17*, 167–190. [CrossRef]
7. Szołtysek, M.; Poniat, R.; Gruber, S. Age heaping patterns in Mosaic data. *Hist. Methods J. Quant. Interdiscip. Hist.* **2018**, *51*, 13–38. [CrossRef]
8. A'Hearn, B.; Baten, J.; Crayen, D. Quantifying Quantitative Literacy: Age Heaping and the History of Human Capital. *J. Econ. Hist.* **2009**, *69*, 783–808. [CrossRef]
9. Can, Y.S.; Kabadayı, M.E. Automatic CNN-Based Arabic Numeral Spotting and Handwritten Digit Recognition by Using Deep Transfer Learning in Ottoman Population Registers. *Appl. Sci.* **2020**, *10*, 5430. [CrossRef]
10. Can, Y.S.; Kabadayı, M.E. CNN-Based Page Segmentation and Object Classification for Counting Population in Ottoman Archival Documentation. *J. Imaging* **2020**, *6*, 32. [CrossRef] [PubMed]
11. Kim, M.S.; Cho, K.T.; Kwag, H.K.; Kim, J.H. Segmentation of handwritten characters for digitalizing Korean historical documents. In *International Workshop on Document Analysis Systems*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 114–124.
12. Zhang, K.; Shen, Z.; Zhou, J.; Dell, M. Information Extraction from Text Regions with Complex Tabular Structure. In *Conference on Neural Information Processing Systems*; Springer: Berlin/Heidelberg, Germany, 2019.
13. Kusetogullari, H.; Yavariabdi, A.; Hall, J.; Lavesson, N. DIGITNET: A Deep Handwritten Digit Detection and Recognition Method Using a New Historical Handwritten Digit Dataset. *Big Data Res.* **2021**, *23*, 100182. [CrossRef]
14. Cheddad, A.; Kusetogullari, H.; Hilmkil, A.; Sundin, L.; Yavariabdi, A.; Aouache, M.; Hall, J. SHIBR—The Swedish Historical Birth Records: A semi-annotated dataset. *Neural Comput. Appl.* **2021**, *1*, 1–13. doi: 10.1007/s00521-021-06207-z. [CrossRef]
15. Khandelwal, A.; Choudhury, P.; Sarkar, R.; Basu, S.; Nasipuri, M.; Das, N. Text line segmentation for unconstrained handwritten document images using neighborhood connected component analysis. In *International Conference on Pattern Recognition and Machine Intelligence*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 369–374.
16. Boussellaa, W.; Zahour, A.; Elabed, H.; Benabdelhafid, A.; Alimi, A.M. Unsupervised block covering analysis for text-line segmentation of Arabic ancient handwritten document images. In Proceedings of the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 1929–1932.
17. Khayyat, M.; Lam, L.; Suen, C.Y.; Yin, F.; Liu, C.L. Arabic handwritten text line extraction by applying an adaptive mask to morphological dilation. In Proceedings of the 2012 10th IAPR International Workshop on Document Analysis Systems, Gold Coast, Australia, 27–29 March 2012; pp. 100–104.
18. Adiguzel, H.; Sahin, E.; Duygulu, P. A hybrid for line segmentation in handwritten documents. In Proceedings of the 2012 International Conference on Frontiers in Handwriting Recognition, Bari, Italy, 18–20 September 2012; pp. 503–508.
19. El-Sawy, A.; Hazem, E.B.; Loey, M. CNN for handwritten Arabic digits recognition based on LeNet-5. In *International Conference on Advanced Intelligent Systems and Informatics*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 566–575.
20. Hoda Dataset. Available online: http://farsiocr.ir (accessed on 30 June 2020).
21. Khosravi, H.; Kabir, E. Introducing a very large dataset of handwritten Farsi digits and a study on their varieties. *Pattern Recognit. Lett.* **2007**, *28*, 1133–1141. [CrossRef]
22. ADBase Dataset. Available online: http://datacenter.aucegypt.edu/shazeem/ (accessed on 30 June 2020).
23. CMATERDB 3.1.3 Dataset. Available online: https://code.google.com/archive/p/cmaterdb/downloads (accessed on 30 October 2020).
24. Dehghanian, A.; Ghods, V. Farsi Handwriting Digit Recognition Based on Convolutional Neural Networks. In Proceedings of the 2018 6th International Symposium on Computational and Business Intelligence (ISCBI), Basel, Switzerland, 27–29 August 2018; pp. 65–68.
25. Ghofrani, A.; Toroghi, R.M. Capsule-Based Persian/Arabic Robust Handwritten Digit Recognition Using EM Routing. In Proceedings of the 2019 4th International Conference on Pattern Recognition and Image Analysis (IPRIA), Tehran, Iran, 6–7 March 2019; pp. 168–172.

26. Farahbakhsh, E.; Kozegar, E.; Soryani, M. Improving Persian digit recognition by combining data augmentation and AlexNet. In Proceedings of the 2017 10th Iranian Conference on Machine Vision and Image Processing (MVIP), Isfahan, Iran, 22–23 November 2017; pp. 265–270.
27. Ahamed, P.; Kundu, S.; Khan, T.; Bhateja, V.; Sarkar, R.; Mollah, A.F. Handwritten Arabic numerals recognition using convolutional neural network. *J. Ambient. Intell. Humaniz. Comput.* **2020**, *11*, 5445–5457. [CrossRef]
28. Alani, A.A. Arabic handwritten digit recognition based on restricted Boltzmann machine and convolutional neural networks. *Information* **2017**, *8*, 142. [CrossRef]
29. Can, Y.S.; Kabadayı, M.E. Curation of Historical Arabic Handwritten Digit Datasets from Ottoman Population Registers: A Deep Transfer Learning Case Study. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020; pp. 1853–1860.
30. Oliveira, S.A.; Seguin, B.; Kaplan, F. dhSegment: A generic deep-learning approach for document segmentation. In Proceedings of the 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), Niagara Falls, NY, USA, 5–8 August 2018; pp. 7–12.
31. Can, Y.S.; Kabadayı, M.E. Line Segmentation of Individual Demographic Data from Arabic Handwritten Population Registers of Ottoman Empire. In Proceedings of the 2021 IEEE 4th International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR), Lausanne, Switzerland, 5–10 September 2021; Springer: Cham, Switzerland; 2021; pp. 312–321. [CrossRef]