Advances in Science and Technology Research Journal

Volume 14, Issue 3, September 2020, pages 30–38
https://doi.org/10.12913/22998624/122567

Received: 2020.05.15
Revised: 2020.06.30
Accepted: 2020.07.15
Available online: 2020.07.25

Development of Extensive Polish Handwritten Characters Database for Text Recognition Research

Mikhail Tokovarov[1], Monika Kaczorowska[1*], Marek Miłosz[1]

[1] Department of Computer Science, Lublin University of Technology, ul. Nadbystrzycka 36B, 20-618, Lublin, Poland

[*] Corresponding author's e-mail: m.kaczorowska@pollub.pl

ABSTRACT

In the modern world, fast and efficient processing of non-digital (handwritten or typed) texts is the task of extreme importance. Similar to many other fields, optical character recognition (OCR) benefits from the application of machine learning (ML) which allows developing effective and accurate methods. In order to achieve good performance, a machine learning algorithm requires great amount of data. Nowadays, a large database of handwritten characters prepared by National Institute of Standards and Technology (NIST), USA, can be used for training an ML model. However, significant differences between the manners of handwriting exist in the US and Poland. That fact, along with the absence of Polish diacritical marks, causes the NIST database to be less useful for development of an OCR model for the Polish language. According to the best of the authors' knowledge, no database with samples of Polish handwriting exists. The present research is focused at filling this gap, i.e. gathering and preparing an extensive database of Polish handwritten characters. The paper presents the very first database of Polish handwriting samples. The database is by far larger than all the datasets used in the previous attempts of implementing OCR for the Polish handwriting. It is also the first fully publicly accessible database of Polish handwriting of this scale. The same method and developed tools can be used to build handwritten characters databases of other languages.

Keywords: OCR, Handwriting character samples, Database for optical character recognition, Polish handwritten characters database

INTRODUCTION

OCR is highly profitable in many fields such as data input, e.g. processing handwritten documents in various government institutions, quicker ways of reading information from official documents such as passport and ID cards on airports, scanning of car number plates on a car park and surveillance system, digitalizing the scientific literature published before the digital era, making it searchable [7, 21]. In order to develop an OCR system, various models can be used. One of the most dynamically developing paradigms is machine learning – the approach based on statistics and optimization. Since it widely applies the methods of statistics, such models require the great amount of data to achieve sufficient performance [4, 12, 16]. Most often, this kind of data is gathered in a form of samples of handwritten text, where the person providing the sample is asked to fill in a form writing some specific text into fields [12, 20, 21]. The forms usually contain the letters, digits and other characters. The letters can be presented separately or be included into words and sentences. The filled forms later undergo the procedure of characters extraction, the extracted characters are pre-processed (e. g: centering, binearization, denoising) [4, 5, 19].

One of the most well-known databases of handwritten characters is the NIST dataset elaborated by National Institute of Standards and Technology [7]. This database contains a large number of handwritten characters gathered in the USA. Thus, it can potentially be used for training a machine learning model for the OCR tasks. However, the science presents the numerous evidences

of diversity of handwriting manners in various countries, which renders the American database less useful for application in Poland [3, 17].

Along with the Latin alphabet, the examples of databases containing the samples of other writing systems can be found in the literature [1, 9, 11]. The said dataset was used for developing the handwriting recognition algorithm [2, 22], which shows that a large publicly available dataset is an essential factor for creating effective handwriting recognition algorithms.

Another reason for creating a Polish database of handwritten characters is the fact that the American database contains only the characters of basic Latin alphabet without special Polish characters having diacritics such as ą, ę, ć, ł, ń, ó, ź, ż. Owing to these facts, the task of creating an extensive database with the samples of Polish handwritten characters appears to be important.

The literature review reveals several attempts of developing an OCR pipeline dedicated to the Polish handwriting [8, 10]. However, they did not involve such extensive handwriting database as the present research does. Grzelak et al. extended the EMNIST dataset with two Polish letters, namely, "Ą" and "Ć" [8]. On the other hand, Kurzyński and Sas analyzed the dataset containing only Polish handwriting, the dataset included the names and the surnames of broad set of medical patients [10]. However, the complete words were used in training a classifier.

Turnbull et al. analyzed the datasets containing the Polish handwriting in order to find the features allowing to distinguish between the Polish and English handwriting; the dataset, however, contained only 53 Polish and 52 English handwriting samples [18].

Górska and Janicki developed a model allowing to recognize the human extraversion level based on handwriting [6]. They examined the handwriting of numerous group including 883 persons: 404 men and 479 women. Although the number of samples was quite high, separate letters were not extracted, instead the author extracted such features as: word spacing, stability of pressure, handwriting regularity etc.

On the basis of the literature review conducted, the authors of the present article can conclude that the database presented in the paper is unique from several points of view:

it is the first large-scale database of Polish handwriting samples, containing over 530 thousands of characters in the forms of separate digits, letters (common Latin and Polish diacritics) and syntax signs;

- it is the largest dataset of handwriting samples collected in Poland;
- it is the first fully labeled and preprocessed dataset of Polish language handwriting samples ready for use in OCR;
- it is the first fully publicly accessible database of Polish handwriting of this scale.

Polish Handwritten Sample Form

The data were gathered with the use of special, one page forms, called Polish Handwritten Sample Form (PHSF). PHSF was elaborated by the authors of the present paper. The forms are anonymous, i.e. the participant did not have to provide the personal data, such as name, surname, etc. Only sex and year of birth were collected for the statistical purposes. Additionally, the information regarding the participants groups was collected using the code form (e.g. 2-B, 2-C). Each form contains the date of filling as well. The PHSF contains 16 fields where every Polish language character appears at least 3 times. The samples of handwriting were collected either in the form of separate letters or sentences. In order to ensure the appearance of a complete alphabet in the sentences, palindromes where chosen. A palindrome is a sentence containing all the letters of an alphabet in particular language. For English language, the following sentence was used: "The quick brown fox jumps over the lazy dog", while for Polish: "Mężny bądź chroń pułk twój i sześć flag" was chosen.

Participants were asked to write the characters with spaces, possibly avoiding crossing the borders of the fields.

On the other side of the form, the well known poetry (called "Invocation") of Adam Mickiewicz was provided. The participants had to write the poetry in the normal way without spaces between each letter. Figure 1 and Figure 2 present the first and the second side of PHSF.

Procedure

Overview

The students of various Polish universities, along with the persons of productive age, were the main part of the participants. The said group was chosen due to the fact that it includes the persons

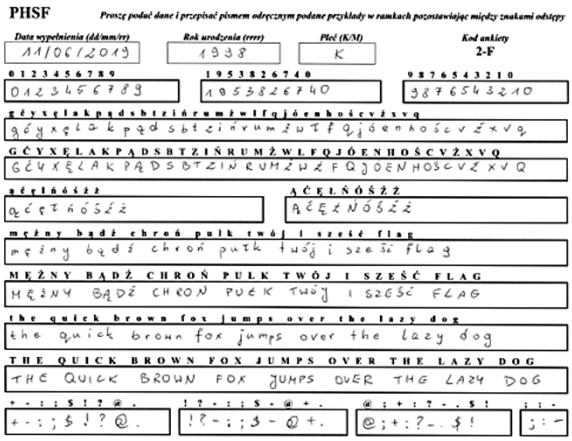

Fig. 1. Filled Polish handwritten sample form (first side)

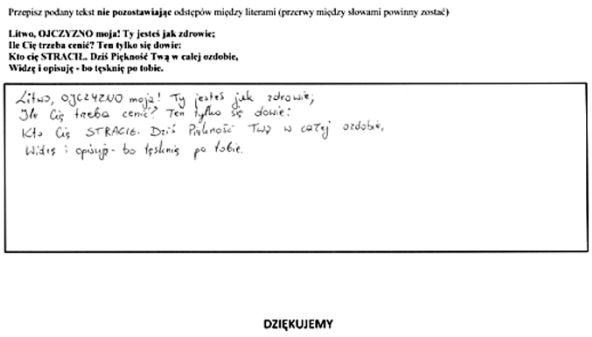

Fig. 2. Filled Polish handwritten sample form (second side)

who either are employed or are coming to the labor market. Moreover, the approach of teaching handwriting is changing nowadays – the subject of formal calligraphy disappears from schools and the handwriting itself changes its character.

The participants from both technical and human science specialties took part in the research. They were asked to fill in the forms. The average time of filling in the form was 10 minutes. Over 2000 completed forms were collected.

Initial selection including the rejection of insufficiently filled forms was carried out by hand. The forms were rejected in the case when the characters were written without spaces and/or when many characters crossed the borders of the fields. The mentioned rejection was necessary due to high complexity of processing such insufficiently filled forms.

The next step was scanning the forms with the 600 dpi resolution. The obtained files were transformed from RGB to grey scale. Afterwards, they were thresholded and the colors were inversed, i.e. black parts became white and white parts became black. Subsequently, the procedure of character extraction was performed.

The application dedicated for character extraction was developed. The Python 3.6 was used for it. The following libraries were used: numpy, opencv, pyQt 5 and Pillow. The application had the following main functionalities:

- for both PHSF sides and separate fields: editing, dilatation, erosion, rejection;
- for processing separate characters: deleting, rejection, joining neighboring characters.

The application has a user friendly graphical interface – Figure 3.

Character extraction

The special procedure was developed to extract of characters. In order to describe it, let us assume that the scanned form is referred to as F being an integer grid, i.e.: $F \in \mathbb{N}^{n \times m}$, where n and m are, correspondingly, its height and width.

The procedure of character extraction was performed in the following steps:

- removal of a form header with the age/sex information not containing the characters for extraction,
- field mask extraction,

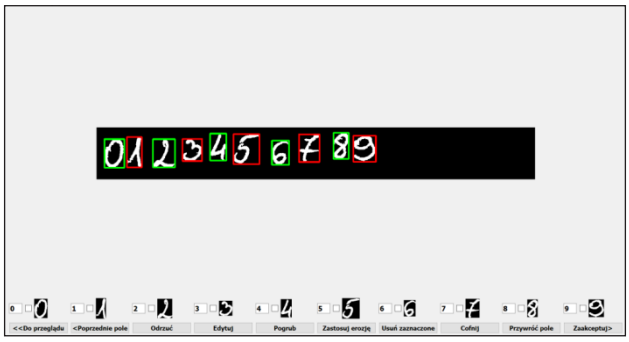

Fig. 3. GUI of the dedicated application

- aligning a scanned form,
- detecting and extracting separate lines of the form,
- detecting and extracting separate fields in each line,
- detecting and extracting separate characters in each field.

The description of the listed steps is presented below.

Removal of a form header is carried out by setting the values of pixels of the upper part of the form to 0. "The upper part" is defined as 18.5% the form height. The value 18.5% is set empirically. In rare cases, when some elements of header are not eliminated, the user of the dedicated application has the option of correcting the form manually.

In order to perform the next steps, i.e. aligning the forms as well as extracting separate lines and fields, the field mask was extracted. Hereinafter, the field mask is referred to as M_f.

The field mask is obtained by sequent application of two filters: vertical and horizontal f_v (and f_m). Each filter is used in the two morphological operations: erosion and dilation [13, 15]. The mentioned filters have the following structure:

$$f_h = [1 \quad 1 \quad \cdots \quad 1]_{1\times p} \tag{1}$$

$$f_v = [1 \quad 1 \quad \cdots \quad 1]^{\mathrm{T}}_{1\times q} \tag{2}$$

The horizontal filter f_h is a row vector of ones of length p and the vertical filter f_v is a column vector of ones of length q. The values of p and q were set empirically in the way ensuring the best performance of the algorithm. Number p was set to 220 and q – to 200.

The above-mentioned operations of erosion and dilation were applied for obtaining a field mask M_f, which was the union of two masks: the horizontal and the vertical (M_h and M_v). The said morphological operations were applied in the binary way, i.e.:

for erosion:

- for the vertical mask: a pixel $M_v[i,j]$ of M_v was set to 0 in case if any pixel of $F[i - q/2{:}\, i + q/2 - 1, j]$ equals to 0, otherwise it was set to 255;
- for the horizontal mask: a pixel $M_h[i,j]$ of M_h was set to 0 in case if any pixel of $F[i, j - p/2{:}\, j + p/2 - 1]$ equals to 0, otherwise it was set to 255;

for dilation:

- for the vertical mask: a pixel $M_v[i,j]$of M_v was set to 255 in case if any pixel of $F[i - q/2{:}\, i + q/2 - 1, j]$ equals to 255, otherwise it was set to 0;
- for the horizontal mask: a pixel $M_h[i,j]$ of M_h was set to 255 in case if any pixel of the subset $F[i, j - p/2{:}\, j + p/2 - 1]$ equals to 255, otherwise it was set to 0.

After obtaining the vertical and horizontal masks (M_h and M_v), the field mask was obtained in the following way:

$$M_{\mathrm{f}}[i,j] = \begin{cases} 0, \; M_{\mathrm{h}}[i,j] = 0 \; and \; M_{\mathrm{v}}[i,j] = 0 \\ 255, otherwise \end{cases} \tag{3}$$

Figure 4 presents the original scanned form and the obtained field mask. As it can be seen , the procedure of field mask extraction allows eliminating noise and makes the further steps of the form processing possible. M_f, M_h and M_v are used in the further processing steps.

The procedure of the form alignment is an essential step due to the fact that the forms are usually placed in the scanner with slight misalignment, which can lead to poorer extraction of the characters. In order to align a form, it should be rotated around its center at the angle α. The angle α is obtained in the following way:

$$\alpha = arctg\left(\frac{y_r - y_l}{x_r - x_l}\right) \tag{4}$$

Where (χ_l, γ_l) are the coordinates of the closest to left-top corner of the F white pixel, while (χ_r, γ_r) are the coordinates of closest to right-top corner of the F white pixel. The term "closest" is used here in the sense or L1 metric. Due to the fact that all the coordinates are positive numbers and any χ is less than the image width, the modulo operation can be omitted. This can be expressed in the following form:

$$(x_l, y_l) = argmin(x + y) \tag{5}$$

$$(x_r, y_r) = argmin(m - x + y) \tag{6}$$

where: $(x, y)\ \epsilon\ C$, C being the set containing all the coordinates of white pixels of M_f.

The Figure 5 presents the schematic explanation of the formula (4).

The direction of rotation is counterclockwise in the case if α is positive and clockwise if α is negative [14]. After the form is aligned, the next step, named extraction of separate lines, is performed. In order to obtain separate lines, horizontal mask, M_h, is analyzed. A set H_h contains the values equal either to 1 or 0 is obtained. The number of values in the set is equal to the number of rows in M_h. The value 1 corresponds to the row containing at least one white pixel, 0 corresponds to the row, where all the pixels are black. As a horizontal filter was applied, M_h contains only horizontal lines. The Figure 6 presents the example of H_h.

After obtaining H_h the differences D_h between every pair of neighboring elements are computed. The example is presented in Figure 7.

On the basis of D_h, the coordinates of vertical borders of the fields can be obtained in the following way: first, two separate sets of order numbers starting from 0 should be assigned to the positive and negative pikes of D_h.The set of borders (b_i^u, b_i^l) for i-th line can be obtained in the following way:

$$b_i^u = d_{2i}^+ \tag{7}$$

$$b_i^l = d_{2i+1}^- \tag{8}$$

where: (b_i^u, b_i^l) – upper and lower borders of i-th line;

a) b)

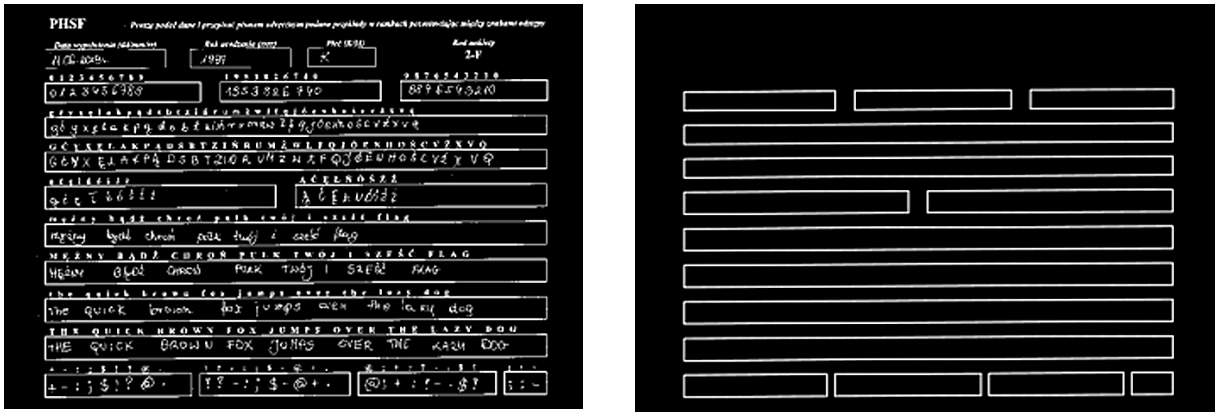

Fig. 4. Scanned sample form (left) and field mask extracted from it (right)

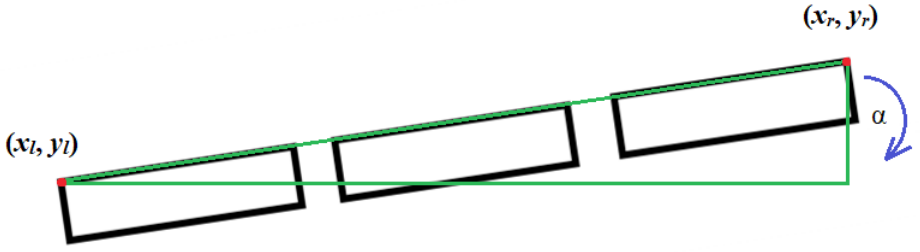

Fig. 5. Handwriting sample form alignment

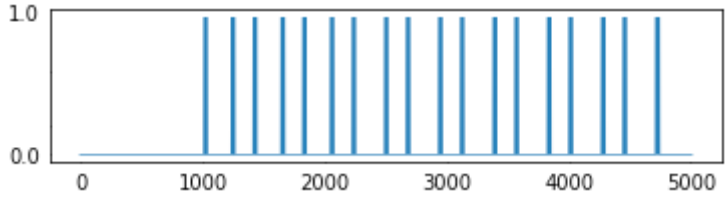

Fig. 6. Example of histogram indicating presence/absence of horizontal line in every row of M_f

d_k^+ – value corresponding to *k*-th positive peak;
d_k^- – value corresponding to *k*-th negative peak.

Having obtained vertical borders, separate fields of the *i*-th line can be extracted by calculating vertical histograms of the lines of M_v, bounded by *i*-th set of borders (b_i^u, b_i^l). The processing of the obtained histograms is performed in the way similar to extracting the vertical line borders, after that a set (χ_l, γ_l), (χ_r, γ_r) i for *i*-th field is obtained.

(χ_l, γ_l) are the coordinates of top-left corner and (χ_r, γ_r) are the coordinates of bottom-right corner of *i*-th field.

In order to extract the characters, M_f was subtracted from element-wise, *F* allowing obtaining the sample form without border boxes.

The character extraction is carried out from the separate fields. A histogram approach is used for that task as well. Figure 8 presents the extracted characters marked with colors. Sometimes, the characters cannot be extracted separately due to various factors, e.g.: joint handwriting, tilted letters, and bad scan quality. In those cases, the user has the possibility of correcting a separate field manually.

DESCRIPTION OF DATABASE

Character images representation

Using the gathered PHSFs and developed application the Polish Handwritten Characters Database (PHCD) was created. One of the PHCD part contains the images with characters extracted from the forms.

Decimal integer numbers (codes) were assigned to the characters in the form in the following way:

- digits (0–9) – 0–9,
- lower case Latin letters (a-z) – 10–35,
- upper case Latin letter (A-Z) – 36–61,
- lower case special Polish letters – 62–70,
- upper case special Polish letters – 71–79,
- selected punctuation marks – 80–88.

The database is structured in the following way: the main directory is named *phsf* (acronym from Polish Handwriting Sample Form). It includes two subdirectories: characters and invocation. Each of these directories contains a *png* subdirectory. The *png* directory located in the characters directory contains directories named with the numbers assigned to the characters. Every file contains the picture with a single character. The *png* directory located in the invocation directory

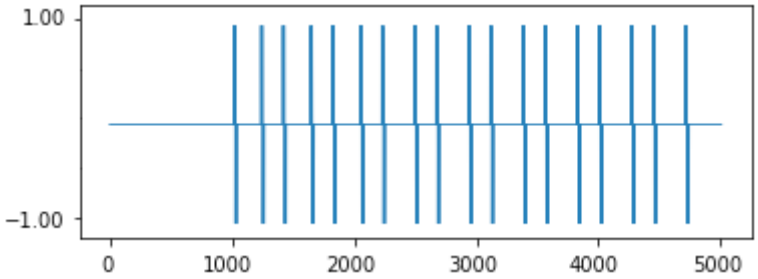

Fig. 7. Differences D_h between each pair of values in h_h

Fig. 8. Example of extracted characters

contains the extracted Invocations (long text – Figure 2). Figure 9 presents the diagram with the *phsf* directory structure.

The names of the files located in the directories 0–88 have the following format:

CharacterNumber_FileNumber_YearOfBirth_Sex_Code.png,

where:

- CharacterNumber – the code assigned in accordance with the rule presented above,
- FileNumber – the order number of the file in the present directory; the file number was written in 4 digit number,
- YearOfBirth – birth year of the participant filling the form; year is presented by 2 digit number,
- Sex – sex of the participant filling the form; sex is presented by K – female and M – male,
- Code – corresponds to the participants groups; the code is formed by 1 digit and 1 letter.

The examples of files name are presented in Figure 9. For example: 47_0000_94_K_1A.png means that the number of character is 47 (character M), file number – 0000, year of birth – 1994, female, code 1A.

The PHCD contains at least 6,000 samples of each character. The total number of samples is about 530,000. Each character image is organized in the following way: the character was centered and scaled into box with the following dimensions: width – 20 px, high – 32 px. This rectangle is centered on the 32px on 32 px square. Figure 10 presents the examples of single character images stored in the PHCD.

Character numerical representation

Additionally, the PHCD contains the *ocr_files* directory where the data preprocessed in a form suitable for application in machine learning are collected. The motivation for this approach was the fact that loading separate png files into RAM is extremely time consuming and requires specific libraries for processing images.

The *ocr_files* directory of PHCD includes four following files:

- signs.npy – contains the signs in a form of 3D numpy array filled with uint8 values, where: 0 – represented black pixels and 255 – white. The dimensions are the following: total number of files-by-32-by-32.
- labels_int.npy – contains 1D numpy array filled with uint8 values. A value on the i-th position in this array corresponds to i-th 32-by-32 array in the signs.npy file.
- dictionary.json – contains the dictionary where the characters are assigned to decimal numbers according to the princip presented in the chapter 4.1.
- binarized_signs.npy – contains 2D array filled with uint8 values. The dimensions of the array are the following: total number of characters-by-128.This file contains the compressed data.

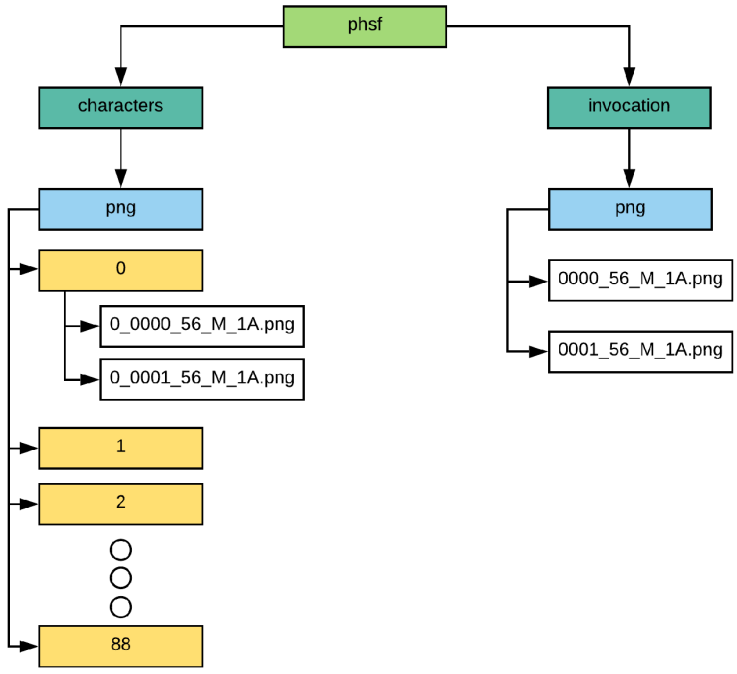

Fig. 9. The *phsf* directory structure

Fig. 10. Examples of a single character

CONCLUSION

The database of Polish handwritten characters has been developed in accordance with the procedure similar to the one applied for collecting the data contained in the NIST characters dataset [7]. The developed database contains the most important characters such as lower case and upper case letters, including the letters with Polish diacritics, digits and syntax characters. The data were gathered from extensive group of participants including students of various specialties such as computer science, electrical engineering, mechanical engineering, civil engineering, economics, logistics, management, mechatronics and mathematics. Wide range of academic specialties ensured diversity of handwriting.

The Python 3.6 programming language was used for developing character extraction application. The following libraries were applied: OpenCV, Pillow, numpy, PyQt5. The handwritten samples were prepared for use in machine learning in the following way: the characters were extracted, de-noised and scaled to specific size. Each character is stored in two ways: saved as a separate image and jointly with the other characters in .npy files. The second approach ensures significantly shorter loading time for the dataset.

The presented PHCD contains the labeled and fully prepared data, which can be used by researchers for developing the models of optical character recognition for the Polish language and other text recognition research. The presented method and developed tools can be used to build handwritten characters databases of other languages. PHCD is publicly available free of charge at http://cs.pollub.pl/phcd.

REFERENCES

1. Bhattacharya, U., Chaudhuri, B. B. Handwritten numeral databases of Indian scripts and multistage recognition of mixed numerals. IEEE transactions on pattern analysis and machine intelligence. 2008, 31(3), 444–457.
2. Dash, K. S., Puhan, N. B., & Panda, G. BESAC: Binary External Symmetry Axis Constellation for unconstrained handwritten character recognition. Pattern Recognition Letters. 2016, 83, 413–422.
3. Dhaka, V. S., Kumar, M., Chaudhary, P. Offline Handwritten English Script Recognition: A Survey. International Journal of Advanced Networking and Applications (IJANA). 2014, 114–124.
4. Garris M. D., Wilkinson R. A. HWSC – Handwritten segmented characters database. In Technical Report Special Database. National Institute of Standards and Technology. 2017.
5. Garris, M. Methods for evaluating the performance of systems intended to recognize characters from image data scanned from forms. 1993
6. Górska, Z., Janicki, A. Recognition of extraversion level based on handwriting and support vector machines. Perceptual and motor skills. 2012, 114(3), 857–869.
7. Grother, P. J. NIST Special Database 19. NIST, Handprinted Forms and Characters Database. National Institute of Standards and Technology. 1995.
8. Grzelak, D., Podlaski, K., Wiatrowski, G. Analyze the effectiveness of an algorithm for identifying Polish characters in handwriting based on neural machine learning technologies. Journal of King Saud University-Computer and Information Sciences. 2019, 1–7.
9. Khosravi, H., Kabir, E. Introducing a very large dataset of handwritten Farsi digits and a study on their varieties. Pattern recognition letters. 2007, 28(10), 1133–1141.
10. Kurzynski, M., Sas, J. Combining Character Level Classifier and Probabilistic Lexicons in Handwritten Word Recognition–Comparative Analysis of Methods. In International Conference on Computer Analysis of Images and Patterns. Springer, Berlin, Heidelberg. 2005, 330–337.
11. Manjusha, K., Kumar, M. A., & Soman, K. P. On developing handwritten character image database for Malayalam language script. Engineering Science and Technology, an International Journal. 2019, 22(2), 637–645.
12. Modi, H., Parikh, M. C. A review on optical character recognition techniques. Int J Comput Appl. 2017, 160(6), 20–24.
13. Pesch, H., Hamdani, M., Forster, J., Ney, H. Analysis of preprocessing techniques for latin handwrit-

ing recognition. In 2012 International Conference on Frontiers in Handwriting Recognition. IEEE. 2012, 280–284.

14. Ravi S., Khan A. M. Morphological operations for image processing: understanding and its applications. 2nd National Conference on VLSI, Signal processing & Communications NCVSComs. 2013.
15. Sachdeva, R., Nagpal, P. Text Localization and Extraction in Images Using Mathematical Morphology and OCR Techniques. International Journal of Scientific Engineering and Research. 2013, 1(1).
16. Shastay, A. Misidentification of Alphanumeric Symbols in Both Handwritten and Computer-Generated Information. Home healthcare now. 2015, 33(6), 338–339.
17. Shinde, A. A., Chougule, D. G. Text Pre-processing and Text Segmentation for OCR. International Journal of Computer Science Engineering and Technology. 2012, 2(1), 810–812.
18. Turnbull, S. J., Jones, A. E., Allen, M. Identification of the class characteristics in the handwriting of Polish people writing in English. Journal of forensic sciences. 2010, 55(5), 1296–1303.
19. Wilkinson R. A., Geist J., Janet S., Grother P. J., Burges C. J., Creecy R., Wilson C. L. The first census optical character recognition system US Department of Commerce, National Institute of Standards and Technology. 1992.
20. Wilkinson, R. A., Garris, M. D., Geist, J. C. Machine-assisted human classification of segmented characters for OCR testing and training. In Character Recognition Technologies. International Society for Optics and Photonics. 1993, 1906, 208–217.
21. Wilson, C. L., Garris M. D. Handprinted character database (HWDB). Technical Report Special Database 1. National Institute of Standards and Technology. 1990.
22. Zarro, Rina D., and Mardin A. Anwer. "Recognition-based online Kurdish character recognition using hidden Markov model and harmony search." Engineering Science and Technology, an International Journal, 2017, 20.2, 783–794.