AHD: English & Persian abnormal handwritten digits dataset

Es'hagh Moutabi
Isfahan University of Technology
Isfahan, Iran
eshaghmoutabi95@gmail.com

Shadrokh Samavi
Isfahan University of Technology
Isfahan, Iran
McMaster University
Hamilton, Canada
samavi@mcmaster.ca

ABSTRACT

In this paper, we present AHD(Abnormal Handwritten Digits), a large dataset of special handwritten digits in English and Persian, that people with various disabilities write. These disabilities have an important impact on the style of writing of people that is most of the time hard to read, and these people are always faced with many problems in education or daily life and work. To tackle these problems smart systems came to use deep learning methods to detect their handwriting, but the lack of a suitable dataset made this progress slow. So we made a huge improvement in creating a proper dataset.

The AHD consists of images collected from several sets of handwritten, using different writing tools, including pencils, pens, markers, etc. Due to the diversity in people and tools, there is a high variance in digits, which results in a more trustworthy dataset. The AHD dataset contains around 3000 digits (from 0 to 9), and with various enhancements like changing contrast, brightness, and adding noises like Gaussian, Poisson, and speckle, the final dataset size is 34000 digits. The entire data collection process is the same for both English and Persian. These automated processes can be used in any language and for any purpose.

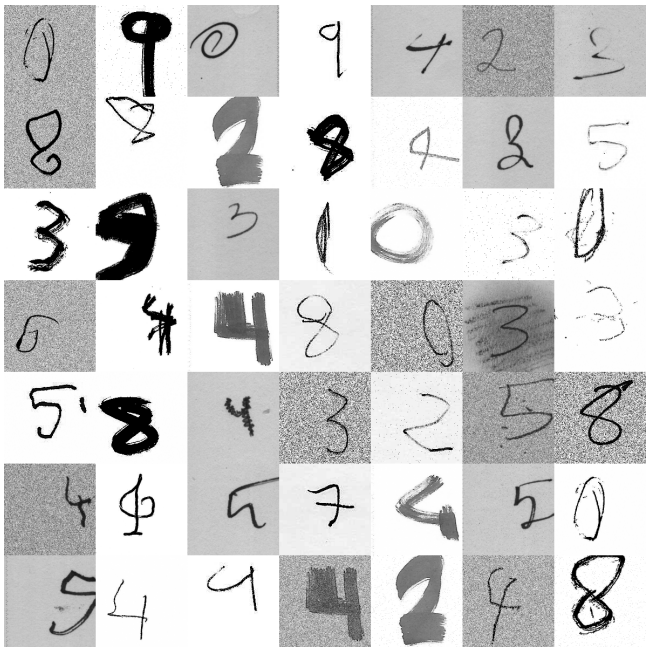

Figure 1: Some samples of the AHD dataset

Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/IzkSensei/HDDS.

1 INTRODUCTION

Research has focused on the problem of detecting numbers and texts in images for many years. In their researches, they tried to solve many challenges, improve their models' accuracy, and review this problem from many aspects. Thus, various datasets and models have been created that try to solve common challenges. [1–3] Not only the high importance of the application of this problem but also, existing many factors and challenges in this problem made this problem interesting for researchers. However, there are still some aspects that have not been studied yet.

Some of these aspects are dependable to many subproblems like lack of proper datasets or high complexity of tasks. Our research aimed to tackle these issues through a study of disability and provided a massive dataset for researchers to improve this field. People with disabilities often have issues with writing that cause their handwriting to be very abnormal, which causes many problems for them in education, work, and daily life, among others. It is challenging to create a dataset due to the wide range of disabilities and, as a result, the various ranges of abnormalities in handwriting. We are primarily focused on keeping a great variety in our dataset so that it is more general and more accurate.

In the rest of this paper, in section 2, we discuss the process of data collection. In section 3, we will explain the dataset and characteristics of digits.

2 DATA COLLECTION

Data collection and its process are crucial because they directly affect the model and its results, and models are susceptible to datasets. Thus, we tried to collect data very precisely with a high degree of sensitivity to details. Because it needs much effort and takes much time, automation is necessary to collect data efficiently. Thus, we decided to design an automated process that will adjust dimensions, filters, and labels of data after collecting and scanning it.

The process of data collection was done in 1 month through several steps. On A4 paper, we used a grid with 10 columns (for 0 to 9) and 15 rows, resulting in squares with the size of 2cm × 2cm, and every digit should be placed in these squares. Each person writes numbers with many writing tools such as a pen, pencil, marker in these cells. Furthermore, some of these tools were pale, which increased the abnormality and diversity of the digits . As a result,

we can have several examples of handwriting from a person with different styles. In order to decrease the impact of bias of models on the resolution of input data, these papers (with the size of A4) were scanned by a scanner with two different resolutions (dpi 100 and 300). A script will automatically cut the grid at specified points, produce images of the dataset, and reshape colourful images to the final dimensions of 28×28 pixels in black and white colour. The original size of cropped cells is 400×400 pixels, and then the script resizes them to 28×28 pixels. So, we have the same number of each digit.

We collected around 3000 images of digits and used augmentation techniques described in section 2 to increase generality. By using these augmentation techniques as the final result, we have 34000 images for English and the same for Persian. In addition, it is possible to use more variation of augmentation in order to increase the size of the dataset

3 DATA CHARACTERISTICS

Various tools used in the process have different effects on the handwriting of people. For having variety as much as we can, we considered some factors in the type of tools which were used:

- Colour: We used many colours like Red, Blue, Black, Orange, Green for some usages where colour is important.
- Thickness: We used many pens, roller pens, and markers of various sizes in order to cover different thicknesses, from a narrow pen with 0.5mm lead to a marker with 0.5cm lead. Thickness has direct effects, especially on people who do not have full control over their movements.
- Paleness: We used writing tools with light or bold ink that left stronger or lighter marks on the paper. Pens that smear ink or are wan increase the level of abnormality in the dataset in some cases where handwriting is old or unreadable.
- Disability: People have physical problems at various levels like injuries to fingers or handicaps or even hand tremors which appear because of injuries to nerves.

In order to cover this factor, we tried to collect data with two approaches:

(1) we tried to collect data from these people as much as we could.
(2) To cover cases where people temporarily have injuries like broken hands that cause them to use their opposite hand, we also asked them to write with the hand that they are not usually used to; for example, a left-handed individual should write with their right hand, and vice versa.

These two main factors helped us collect several high quality samples that contain both diversity and abnormality. As we mentioned earlier, we use several augmentation techniques to make the dataset better in generalization to cover a wide range of characteristics. We considered these items:

(1) Gaussian-distributed additive noise with some random variances (here, we considered 0.1 and 0.3 for the Variance value).
(2) Poisson-distributed noise generated from the data with some random variances (here, we considered 0.5 and 0.4 for the Variance value).
(3) Replaces random pixels with 0 or 1.
(4) Multiplicative noise using $out = image + n \times image$, where n is uniform noise with specified mean variance.

As a result, we generated around 34000 images of handwriting digits. In what follows, we demonstrate the result of each effect and also show some specifications and samples of the dataset.

4 RESULTS AND SAMPLES

In this section, we will show some samples of the dataset and previous filters.

The fig. 2, illustrates three samples of digits from 0 to 9 that each person writes in a row. The fig. 3. shows the augmentations explained earlier, which apply to single digits. The fig. 4 demonstrates the effects of writing tools, such as pen, marker, and so on, as well as some other effects and noises that appear in scanning at different dpi levels.

REFERENCES

[1] Savita Ahlawat, Amit Choudhary, Anand Nayyar, Saurabh Singh, and Byungun Yoon. 2020. Improved handwritten digit recognition using convolutional neural networks (CNN). *Sensors* 20, 12 (2020), 3344.
[2] Hüseyin Kusetogullari, Amir Yavariabdi, Abbas Cheddad, Håkan Grahn, and Hall Johan. 2020. Ardis: a swedish historical handwritten digit dataset. *Neural computing & applications (Print)* 32, 21 (2020), 16505–16518.
[3] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

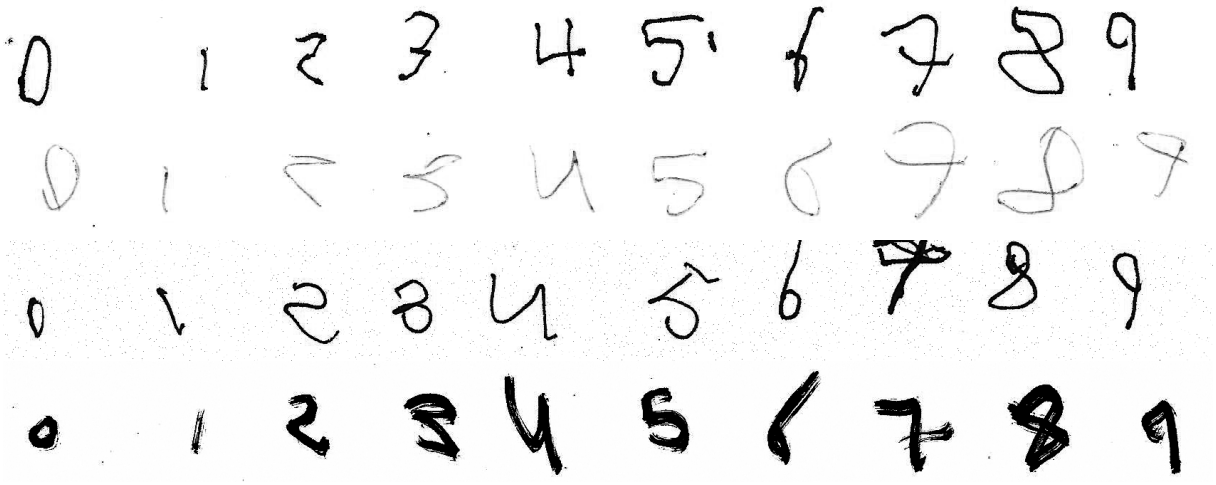

Figure 2: Some set of handwriting digits

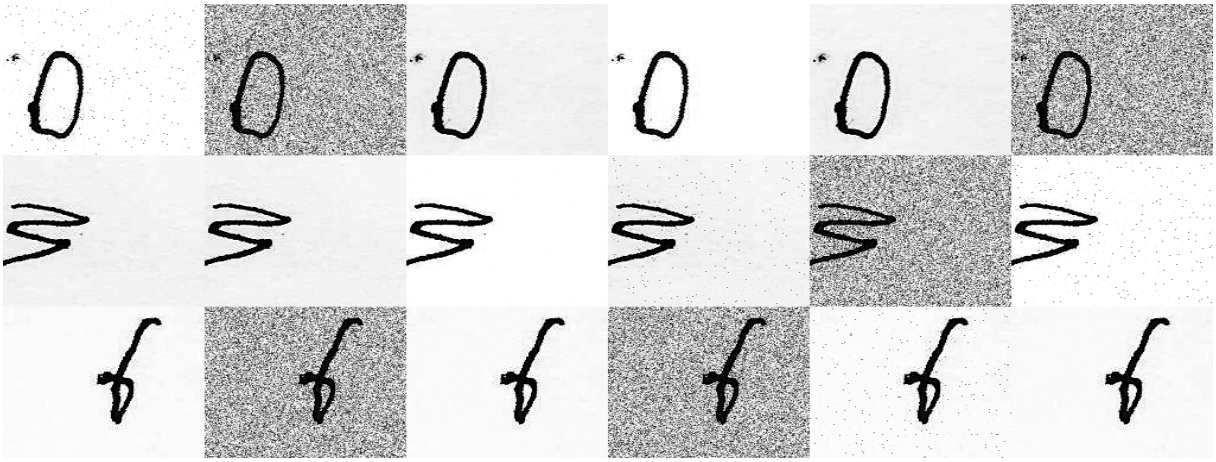

Figure 3: Various augmentations that applied on each sample

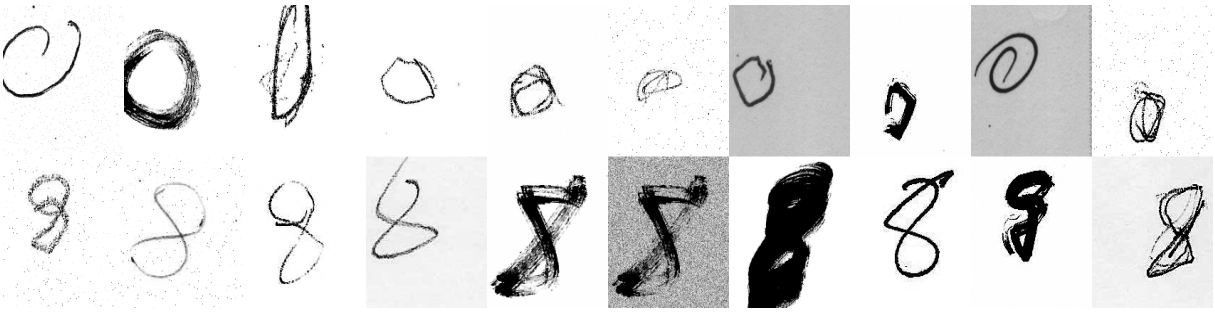

Figure 4: Effects of writing tools and scanning option on digits