# ARDIS

The Swedish Dataset of Historical Handwritten Digits

Department of Computer Science, Blekinge Tekniska Högskola, SE-371 79, Karlskrona, Sweden.

I. Description of the Data Sets

This is a new image-based handwritten historical digit dataset named ARDIS (Arkiv Digital Sweden). The images in ARDIS dataset are extracted from 15.000 Swedish church records which were written by different priests with various handwriting styles in the nineteenth and twentieth centuries. The constructed dataset consists of three single digit datasets and one digit strings dataset. The digit strings dataset includes 10.000 samples in Red-Green-Blue (RGB) color space, whereas, the other datasets contain 7.600 single digit images in different color spaces. Figure 1 illustrates handwritten digit images from different datasets in ARDIS.

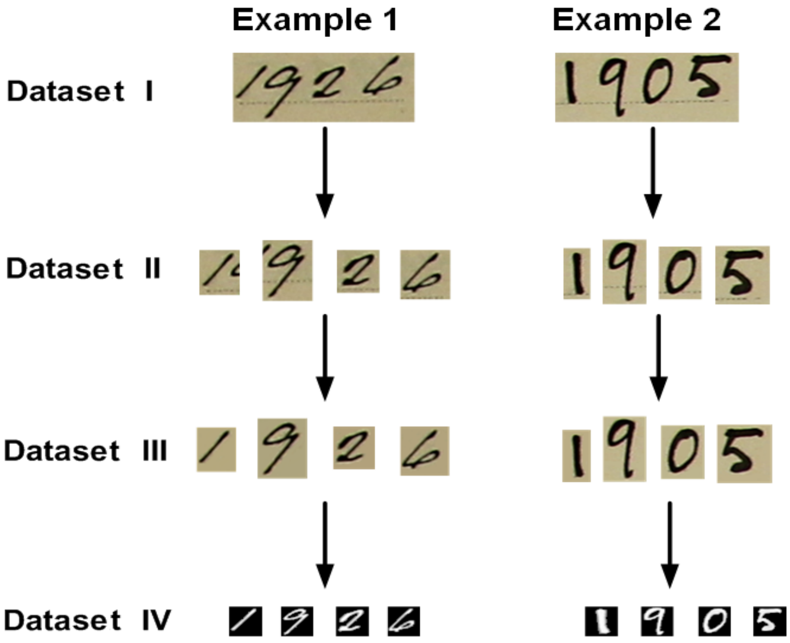

Figure 1. Examples of handwritten digit images from the different datasets in ARDIS.

II. Use of the Materials

If you use any of these data sets, please cite that as:

Huseyin Kusetogullari, Amir Yavariabdi, Abbas Cheddad, Håkan Grahn and Johan Hall, 2019, "ARDIS: A Swedish Historical Handwritten Digit Dataset," Neural Computing and Applications, Springer. DOI: 10.1007/s00521-019-04163-3

Link to the paper, Click here

III. Download Links

#### ARDIS DATASET_I:
This date string image data set contains 10000 images of four digit characters and is divided into the following three parts cropped

automatically from the original full document images (more info here >> Readme.pdf <<) :

1. Part I: This set contains 3977 RGB images in JPG format.
2. Part II: This set contains 4503 RGB images in JPG format.
3. Part III: This set contains 1520 RGB images in JPG format.

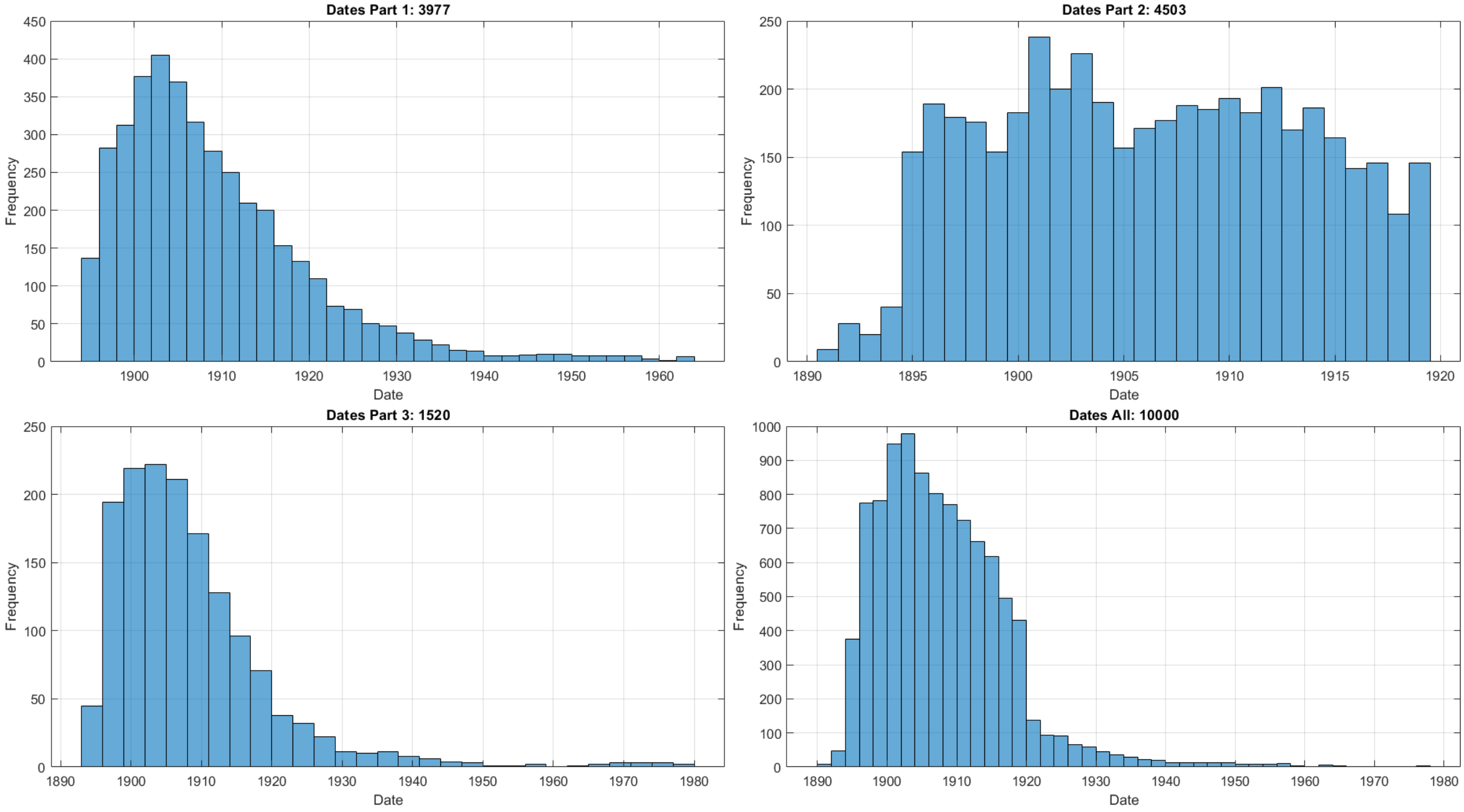

Figure 2. Date Distribution.

#### ARDIS DATASET_II:
This dataset contains 7600 corrupted and noisy handwritten digit images. You can use 6600 images for training and 1000 for testing.

ARDIS_DATASET_II download link: Click here

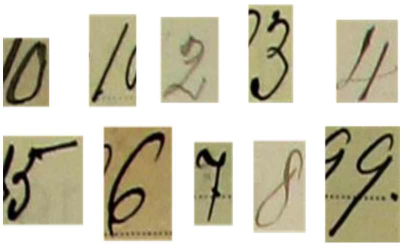

Figure 3. Corrupted Handwritten Digit Images.

#### ARDIS DATASET_III:
This dataset contains 7600 handwritten digit images with clean background. You can use 6600 images for training and 1000 for testing.

ARDIS_DATASET_III download link: Click here

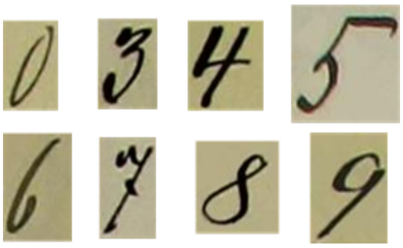

Figure 4. Handwritten Digit Images.

#### ARDIS DATASET_IV:
This dataset contains 6600 training and 1000 testing images in .csv files. The digit images in this dataset are same format with the MNIST and the USPS digit image datasets. The results of different machine learning methods in our accepted paper show that the ARDIS dataset is different

than the MNIST and the USPS datasets.

1. ARDIS_train_2828.csv
2. ARDIS_train_labels.csv
3. ARDIS_test_2828.csv
4. ARDIS_test_labels.csv

ARDIS_DATASET_IV download link: Click here

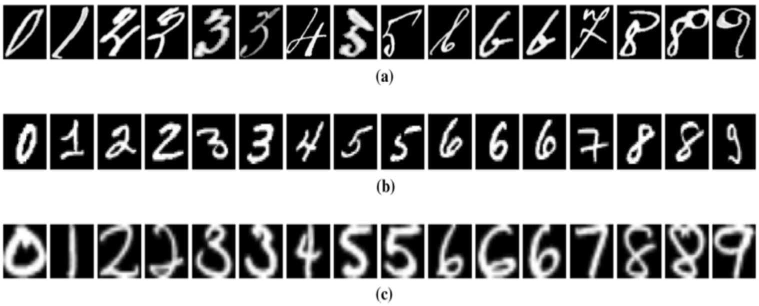

Figure 5. Illustration of digit values from 0 to 9: a) ARDIS, b) MNIST, and c) USPS

IV. Implementation

#### DATASET_IV
#### In Python

```
x_train=np.loadtxt('.../ARDIS_train_2828.csv', dtype='float')
x_test=np.loadtxt('.../ARDIS_test_2828.csv', dtype='float')
y_train=np.loadtxt('.../ARDIS_train_labels.csv', dtype='float')
y_test=np.loadtxt('.../ARDIS_test_labels.csv', dtype='float')
```

#### reshape to be [samples][pixels][width][height]

```
x_train = x_train.reshape(x_train.shape[0], 1, 28, 28).astype('float32')
x_test = x_test.reshape(x_test.shape[0], 1, 28, 28).astype('float32')
```

V. Feedback or Comments

We will be pleased to get your feedback/suggestions to improve the dataset.

✉ Huseyin Kusetogullari
huseyinkusetogullari@gmail.com

Abbas Cheddad
abbas.cheddad@bth.se

Karlskrona, Sweden on: 2019-04-02

Blekinge Institute of Technology