Investigation of Handwriting General Features in Case of Neurological Diseases

Elena Kalcheva[1], Georgi Gluhchev[2]

Abstract – **A method for objective assessment of some neurological diseases (ND) is developed. It is based on the automatic processing and analysis of images of a handwritten script belonging to patients with ND and healthy individuals. In this method connected components are extracted, filtered and analyzed both for the whole script and for each of the automatically separated text-lines. Eight common geometric features describing size, shape and orientation of the handwriting are computed and statistically investigated. The vertical stroke size, aspect ratio and mean slope's angle revealed differences between ND-patients and those from the control group. Also the average stroke width and average stroke number relative to the row length can be used to separate the clinical Parkinsonians from the others. Nevertheless additional features and data are required to obtain reliable evaluation of the handwriting changes due to neurological diseases.**

Keywords – **Neurological diseases, Document image processing, Feature selection, Discriminant ability**

I. Introduction

The disturbances and dysfunction in the human motor control system often represent certain indications of neurological diseases (ND). It is well known that the early diagnosis of many diseases assures a good therapeutic effect in most cases.

In [3] the authors propose several instruments for identification of patients at risk and at early stages of the Parkinson's disease (PD). They suggest some modern invasive techniques, like PET and SPECT studies of the dopaminergic system, analysis of transcranial ultrasonic images, MBIG scintigraphy etc., for recognizing the early stages of the PD. Also, some tests for evaluation of motor abnormalities concerning the velocity, reaction time and precision of the movements are discussed.

The exact study of the effects of the pharmacotherapy on the motor behavior (or so called 'pharmacodynamics') is also very important to determine the appropriate treatment (the amount, timing and spacing of doses) for each patient according to the disease progression and his daily-life requirements [6]. Therefore a simple non-invasive tool for evaluation of the pharmacodynamics for the individual patient is necessary.

In this regard, knowing that motor abnormalities influence the ability to control the movements of the hands, a fine analysis of the handwriting may be useful for objective assessment of the ND. There are some reports on applying handwriting-based techniques to appraise movement disorders. Some of the approaches use only standard neurophysiology tests [5,10]. In [4] the authors combine the Test of Motor Impairments with a SPECT-analysis to examine the possible correlation between long-time treatment and the kinematic parameters of handwriting defects in patients with Wilson's disease.

Moreover, the recent data suggest that handwriting may reveal significant changes many years before the clinical onset of the ND.

In this paper we present a method for objective assessment of some neurological diseases. It is based on the automatic processing and analysis of images of handwritten script belonging to patients with ND and healthy individuals.

II. Automatic Script Processing

In our experiments we use Handwriting Test Form to collect scripts. The scripts consist of a printed sentence in Bulgarian language and the writers are asked to write it in the blank zone below without any requirements about the interpretation. The documents are scanned at 300 dpi and stored as gray-level images.

A. Preprocessing

Most of the algorithms of document analysis and recognition use pre-processing stages to enhance the images. That facilitates the further extraction of relevant information and also increases the accuracy of the automatic measurements. The presence of possible random noise in the scanned images may impede the automatic processing. Therefore in our examinations a 3×3 low-pass filter is applied to remove the noise, consisting of isolated small groups of dark pixels.

At the next step we convert the gray-level images to binary ones using an adaptive thresholding technique [16]. The method is noise resistant and does not depend on the object's area and shape.

B. Connected Components Labeling

Detection of Connected Components (CCs), or blobs, between pixels in binary images is a fundamental step in segmentation of objects and regions within the image. A unique value is assigned to each of them, which allows to separate CC from other blobs.

A CC is denoted as a set of black pixels where each pixel has at least one black 8-neighbor. The original algorithm for

[1]Elena Kalcheva is with the IIT-BAS, Acad. G. Bonchev str., Bl29A, 1113 Sofia, Bulgaria, E-mail: elena@iinf.bas.bg

[2]Georgi Gluhchev is with the IIT-BAS, Acad. G. Bonchev str., Bl29A, 1113 Sofia, Bulgaria, E-mail: gluhchev@iinf.bas.bg

CCs extraction and labeling was developed by Rosenfeld and Pfaltz in 1966 [14]. It performs two passes through the binary image. In the first pass the image is processed from left to right and top to bottom to generate the labels for each pixel and all the equivalent labels are stored in a pair of arrays. In the second pass each label is replaced by the label assigned to its equivalence class. Several papers describe the modifications of the algorithm where the authors try to solve the problem with large images at the second pass where the equivalence arrays become unacceptably huge [9,13]. Another group of algorithms generates Bounding Boxes (BBs) of the Connected Components using the connectivity between the segments of black pixels. Here the binary image is scanned line by line and the arrays of black segments are composed. Then the connectivity between these black segments is examined for each pair of adjacent scanlines. BBs are extended to include all black pixels connected to the previous scanline [2].

In our examinations a simple procedure for Connected Components labeling is used. The binary image is scanned from left to right by rows. The CCs of each row are detected and labeled consecutively using 8-connectivity. The background is assigned 0 and never changes since it is not associated with any other region. Then the equivalent components of adjacent rows are merged and the image is relabeled to obtain a consecutive set of labels for the CCs. The procedure is fast and does not require much space in memory.

A data structure is appointed to each CC. It holds the number of black pixels, the coordinates of the corresponding Bounding Box, the width and height of the CC, the coordinates of its mass center etc. This structure facilitates the manipulation of the components as objects and their further analysis.

After finding all the CCs in the script unusually small and too large components are filtered out. To remove small objects that bring little information we use an absolute criteria – all the components with height or width less than 3 pixels ($\boldsymbol{h}_i < \boldsymbol{3}$ or $\boldsymbol{w}_i < \boldsymbol{3}$, $\boldsymbol{i=1,N}_{cc}$) are eliminated. Then we filter out the CCs with a height greater than twice the mean components height ($\boldsymbol{h}_i > \boldsymbol{2h}_i\_\boldsymbol{av}$) which are considered as 'big' since they are connected across more then one text line. Thus, in our further examinations we deal with the remaining CCs.

C. Text-Lines Segmentation

A text has a linear structure and the physical components corresponding to this linear structure are the text-lines. There are a lot of methods that are successfully applied to the detection of text-lines in printed documents [7,17]. But the layout variability of handwritten materials makes it difficult to obtain a reliable baseline for each of the text-lines. Two basic groups of methods for segmentation of the scripts are presented in the literature – the projection profile techniques [11,12,15,18] and the methods that use CCs analysis [1,8].

The first group of methods is based on the analysis of the regularity of peaks and valleys in the horizontal projection profile of the image. These peaks and valleys correspond respectively to the text-lines and the spaces between them. But the presence of large skew angles (above $\pm 5^\circ$) and overlapping regions (where lines are not parallel) can significantly deteriorate the segmentation results. So the authors often perform prior normalization or use some additional assumptions about the text.

The second group of methods is based on the extraction and clustering of the Connected Components or their corresponding Bounding Boxes. Some authors use a nearest-neighbor technique to merge adjacent components from a same line. Also, Hough transform applied over the set of CCs is used to detect fluctuating baselines and sloped remarks between them [8]. Other authors apply a graph representation where the CCs/BBs are the vertices of the graph, which edges are the lines joining each pair of objects. A Minimum-spanning tree for such graph is computed and the text is segmented using split or merge operations over the branches of the graph according to different rules [1].

Our approach to separate text-line is a simple combination of projection profile technique and CCs analysis. Since the CCs have been already extracted and filtered we compute the horizontal projection profile of their gravity centers. Then the histogram is smoothed and the gaps in the projection profile are taken as rough separators between the lines. The CCs which centers of gravity fall between these borders are attached to the corresponding line. The next step is to obtain the reliable baseline for each of the sets of separated CCs. The straight line which best fits to a given set of points is the regression line determined via the least-squares technique that minimizes the fitting error. Coefficients ***a*** and ***b*** of the line:

$$y = a + bx \tag{1}$$

are computed as follows:

$$b = \frac{N\sum\limits_{i=1}^{N} x_i y_i - \left(\sum\limits_{i=1}^{N} x_i\right)\left(\sum\limits_{i=1}^{N} y_i\right)}{N\sum\limits_{i=1}^{N} x_i^2 - \left(\sum\limits_{i=1}^{N} x_i\right)^2} \tag{2}$$

$$a = \frac{1}{N}\left(\sum_{i=1}^{N} y_i - \sum_{i=1}^{N} x_i\right) \tag{3}$$

where (x_i, y_i), $i = 1, N$ is a point from the set of N points.

The original image of handwritten script and the result image after preprocessing and baseline detection steps are shown on Fig. 1a, b.

III. General Features of Handwritten Text

The preprocessed images of the handwritten scripts are used as input data for extraction of general features at both line and page level.

At the line level we compute several common geometrical parameters that describe size, orientation, shape and baseline deviations of the handwriting. The data structure assigned to each of the CCs at the preprocessing stage facilitates the extraction of some size-related characteristics. We use the height ($\boldsymbol{h}_i$) and the width ($\boldsymbol{w}_i$) of the i^{-th} CC to form the ***aspect ratio***:

$$ar_i = \frac{w_i}{h_i}\ , \tag{4}$$

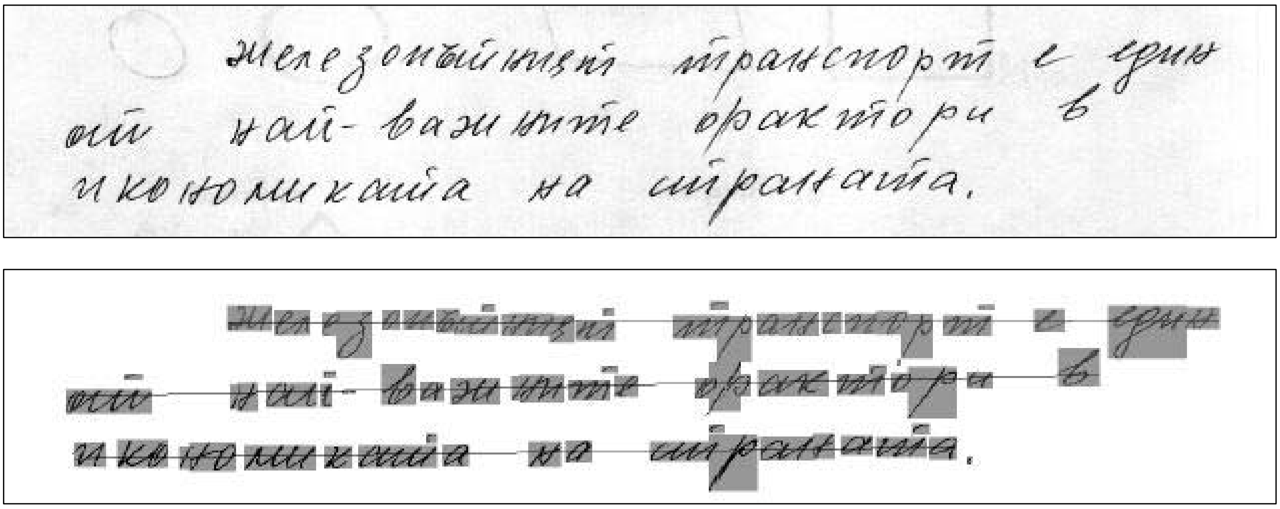

Fig. 1. The original image (a) and the results from the connected components and baselines detection (b).

where ***i*** is the number of the corresponding CC.

The larger ratio the higher probability that a few characters are included in the component. Thus, the ***aspect ratio*** directly gives information about the consistency of the handwriting.

Another feature that fits to the characteristics of different handwritings is the ***stroke to area ratio***. It is defined as:

$$sar_i = \frac{n_i}{w_i h_i} \ , \tag{5}$$

where $\boldsymbol{n}_i$ is the number of the black pixel in the i^{-th} CC. The ***stroke to area ratio*** describes the density of the black pixels within the zone of CC.

These four parameters - $\boldsymbol{h}_i$, $\boldsymbol{w}_i$, $\boldsymbol{ar}_i$ and $\boldsymbol{sar}_i$ are averaged over the number of the CCs ($\boldsymbol{N}_l$) of each text-line (***l***) and the obtained characteristics $\boldsymbol{w\_av}_l$, $\boldsymbol{h\_av}_l$, $\boldsymbol{ar\_av}_l$ and $\boldsymbol{sar\_av}_l$ are used in our examinations. Also, the ***stroke frequency in a row*** is computed as the number of CCs relative to the length of the corresponding baseline:

$$sf_l = \frac{N_l}{Lenght_l} \tag{6}$$

$\boldsymbol{Lenght}_l$ represents the Euclidean distance between the leftmost and the rightmost black pixel of the CCs in the line.

Additional information about the handwritten scripts gives the ***slope angle*** of each text row. Once the baselines of text are obtained it is easy to calculate each slope angle as:

$$slope_l = \arctan(b_l) \ , \tag{7}$$

where b_l is the coefficient from the equation of the regression line Eq. 1.

A certain measure for baseline variability can be defined as:

$$f_l = \frac{S_l}{Lenght_l} \ , \tag{8}$$

where:

$$S_l = \frac{\sum_{i=1}^{N_l} |a_l + b_l x_i - y_i|}{\sqrt{b_l^2 + 1}} \tag{9}$$

is the square root of the residual sum for each regression line and represents the sum of the deviations of the gravity centers of the CCs from the corresponding baseline.

At the page level, all of these parameters are normalized by the number ***R*** of the text-rows in each document. (Same notations but with the capital letters are used for the parameters of the whole script). Also, the average distance between lines is computed as:

$$D\_av = \frac{\sum_{l=1}^{R-1} (a_{l+1} - a_l)}{R-1} \tag{10}$$

Here a_l is the vertical intercept of the beginning of the corresponding baseline.

IV. Experiments and Results

To evaluate the influence of the ND on the handwriting general characteristics several experiments have been carried out. We use the script materials collected from 7 Parkinsonians (Group A), 18 patients with other neurological diseases (Group B) and 10 healthy individuals (Control group C).

The discriminant ability of eight common features of handwriting at line and page level is examined in three experiments:

Experiment 1: A comparison between Control group and the whole group of patients with neurological diseases.

Experiment 2: A comparison between Control group and the group of patients with Parkinson's disease.

Experiment 3: A comparison between the group of Parkinsonians and the group of patients with other neurological diseases.

The two-sided t-criterion is used to estimate the significance of every feature in the experiments. We assume a normal distribution with unknown values of the mean and standard deviation for each parameter in the groups.

The results from the experiments include the evaluation of the minimal and maximal values of the corresponding parameter, its mean and standard deviation, and t-statistics value.

The features of the handwritten scripts that have the greatest values of t are summarized in Table 1. At the page level, vertical stroke size (H), stroke area ratio (*SAR*) and mean slope's angle (*Slope*) reveal differences between the ND-patients and the healthy persons. Moreover, the *SAR* and F parameters show a relatively high t-value in the Experiment 3 where the PD-patients are compared to the other ND-patients.

The tremor, due to certain ND influences the writing control and especially at the beginning and the end of the process. That is confirmed by the line-level analysis where the corresponding features (ar_av and f) have comparatively large t-values for both the first and the last text-lines. Also the slope angle in the first row proves to be a significant feature in the comparison between the ND-patients and the healthy individuals.

Table 1. Summarized results from the experiments

Experiments	Line level features			Page level features
	Line 1	Line 2	Line 3	
Experiment 1 (AB/C)	f, slope, h_av, ar_av	h_av	f, sf, ar_av	H, SAR
Experiment 2 (A/C)	f, ar_av, slope, h_av	ar_av	sf, ar_av, f, sar_av	SAR, Slope
Experiment 3 (A/B)	f, ar_av	ar_av, f, sar_av, sf, w_av	sf ,sar_av, slope,w_av ar_av	SAR, F

V. Conclusion

The motor abnormalities due to neurological diseases influence the ability of patients to control the precise hand movements. Thus, a fine analysis of handwriting could be used as a reliable tool for assessment and evaluation of the ND.

A method for automatic processing of handwriting is proposed to extract general features of handwritten scripts belonging to patients with ND and healthy persons. The statistical significance of 8 parameters is examined at both page and line level using t-test. Most of the features show relatively high discriminant ability in the three implemented experiments.

Nevertheless, the future work must be concentrated on the inner-level analysis of handwritten materials, concerning words and characters, to evaluate the changes in the particular characteristics of handwriting due to neurological diseases.

References

[1] Abuhaiba I. S. I., S. Datta, M. J. J. Holt, *Line extraction and stroke ordering of text pages*, Proc. of the Third Int. Conference on Document Analysis and Recognition (ICDAR), pp. 390-393, Montreal, Canada, August 1995

[2] Amin A., St. Fischer, T. Parkinson, R. Shiu, *Fast algorithm for skew detection*, In Proc. of SPIE'96, 1996

[3] Becker G. et al., *Early diagnosis of Parkinson's disease*, Journal of Neurology, Volume 249, Issue s03, pp iii40-iii48, 2002

[4] Hermann W. et al., *Correlation between automated writing movements and striatal dopaminergic innervation in patients with Wilson's disease*, Journal of Neurology, Volume 249, Issue 8, pp 1082-1087, 2002

[5] Holmen K., K.Ericsson, Y.Forsell, B.Winblad, *Human-Figure-Drawing (HFD) in the screening of different types of dementia in old age*, In Proc. of Sixth Int. Conf. on Handwriting and Drawing, pp. 264-266, Paris, 1993

[6] Jose L.Contreras-Vidal, *Towards non-invasive diagnostic tools and biological markers for Parkinson's disease: A commentary*, Bulletin of the International Graphonomics Society, BIGS, Volume 14, N: 1, pp. 4-5, 2000

[7] Liang J., I. Philips, R. Haralick, *A statistically based, highly accurate text-line segmentation method*, In Proc. of Fifth Int. Conference on Document Analysis and Recognition (ICDAR), pp. 551-554, Bangalore, India, Sept. 20-22, 1999,

[8] Likforman-Sulem L., A. Hanimyan, C. Faure, *A Hough based algorithm for extracting text lines in handwritten documents*, In Proc. of the Third Int. Conference on Document analysis and recognition (ICDAR), pp. 774-777, Montreal, Canada, August 1995

[9] Lumia R., *A new connected components algorithm for virtual memory computers*, IEEE Trans. on CVGIP, Vol. 23, pp. 287-300, 1983

[10] Mai N., C.Marquardt, *Analysis of handwriting movements in brain damaged patients*, In Proc. of Sixth Int. Conf. on Handwriting and Drawing, pp. 240-242, Paris, 1993

[11] Marti U., H. Bunke, *Handwritten Sentence Recognition*, In Proc. of the 15^{th} Int. Conf. on Pattern Recognition (ICPR), Volume 3, pp. 467-470, Barcelona, Spain, 2000

[12] Marti U., H. Bunke, *A full English sentence database for off-line handwriting recognition*, In Proc. of the Fifth Int. Conf. on Document Analysis and Recognition (ICDAR'99), pp. 705 – 708, Bangalore, 1999

[13] Park J.-M, C. Looney, H.-C. Chen, *Fast connected component labeling algorithm using a divide and conquer technique*, In Proc. of the 15^{th} Int. Conference on Computers and Their Applications (CATA), pp 373-376, New Orleans, Louisiana, USA, March 29-31, 2000

[14] Rosenfeld A., J.L. Pfaltz, *Sequential operations in digital processing*, JACM, Vol.13, pp. 471-494, 1966

[15] Said H.E.S., G. Peake, T. Tan, K. Baker, *Writer identification from non-uniformly skewed handwriting images*, On-line Proc. of the Ninth British Machine Vision Conference, 1998

[16] Shapiro V.A., P.K. Veleva, V.S.Sgurev, *An adaptive method for image thresholding*, Proc. of 11th Int. Conf. on Pattern Recognition, pp. 696-699, Hague, 1992

[17] Waked B., S. Bergler, C. Y. Suen, S. Khoury, *Skew detection, page segmentation and script classification of printed document images*, IEEE Int. Conf. on SMC (SMC'98), pp. 4470-4475, San Diego, California, Oct. 1998

[18] Wienecke M., G. A. Fink, G. Sagerer, *Experiments in unconstrained off-line handwritten text recognition*, In Proc. of the 8th Int. Workshop on Frontiers in Handwriting Recognition (IWFHR), Ontario, Canada, August 2002