THE INVOLVEMENT OF INSTITUTE FOR INFORMATION TECHNOLOGIES IN TEXT PROCESSING

Georgi Gluhchev

Abstract: The activities of the Institute of Information Technologies in the area of automatic text processing are outlined. Major problems related to different steps of processing are pointed out together with the shortcomings of the existing solutions.

Keywords: Image Processing, Image Enhancement, Text Segmentation

Introduction

The Institute of Information Technologies was created in 1994 as a successor of the former Institute of Technical Cybernetics, Institute of Engineering Cybernetics and Robotics and Institute of Informatics. Thus, it inherited the scientific traditions and investigations in modern areas like AI, Decision Support Systems, Multicriteria Analysis, Image Processing and Pattern Recognition, Information Processes, Systems and Media, Intelligent Systems and Soft Computing.

The problem of automatic text and handwriting processing is of great importance because its satisfactory solution will allow digitizing millions of printed and handwritten materials all over the world, thus making them broadly available. For example the problem for reliable and secure preservation of the cultural heritage is especially important and urgent one. This is why so many researchers all over the world have been involved in during the last decades.

Figure 1 represents the general scheme of an automated image processing system.

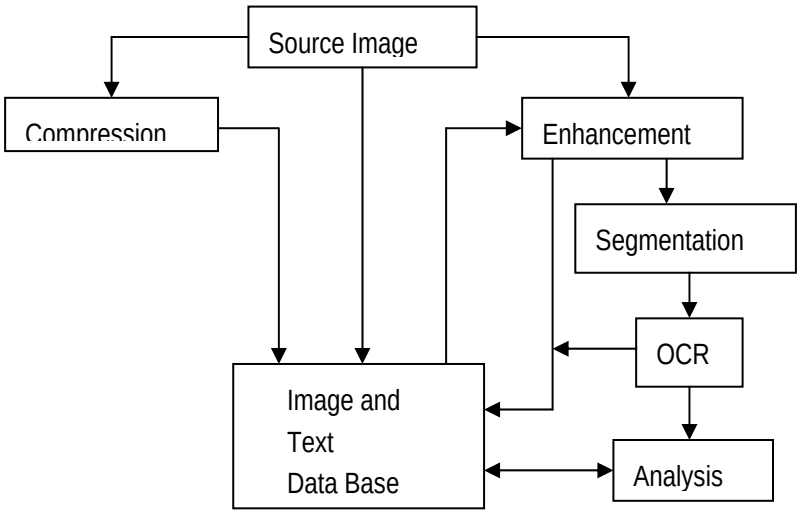

Fig. 1. General text image processing scheme

For more than two decades investigation work carried out by the Laboratory of Image Processing and Pattern Recognition has been aimed at the automatic processing and analysis of handwritten and printed text including letter recognition and writer identification. This led to the development and implementation of computerized systems for similarity estimation of handwritings mainly for forensic purposes. The Lab's research activities and application orientation are summarized in the following table.

Research areas	Application areas
1. Image processing	Handwriting analysis and writer recognition
Contrast improvement	Face recognition
Smoothed image restoration	Speaker recognition
Image segmentation	Moving objects tracking
Automatic and interactive feature extraction	Multimedia applications
Selection of optimal subset of features	Fast and reliable search in large data bases of images
2. Pattern Recognition	
Linear classifiers	Recognition of car license plates
Statistical decision rules	Printed characters recognition
Neural networks	NN for robust control
Clustering	Medical imaging

Image Enhancement

Very often documents that have to be processed are of poor quality due to different factors. This holds especially for ancient manuscripts or printed texts where time and improper conditions (dry or humid air) or handling may cause severe damages, resulting in presence of random and structured noise and diminishing image contrast (Fig.2a). To make the image more pleasing visually, on the one hand, and suitable for further processing on the other, special image quality enhancement techniques have to be applied. Three groups of approaches aimed at noise reduction, contrast enhancement and line refinement could be outlined [Gluhchev G.][Pratt W. K.] . However it must be pointed out that improving image in one aspect may cause its deterioration in another. For example, noise reduction will diminish the image contrast and blur edges and vice–versa.

1. Noise reduction deals with different type of random or structured noise. To diminish the effect of random noise, variances of averaging or median–based filters are used.

2. Contrast improvement is aimed at the increase in color difference between the background and printed or written symbols. The corresponding theory includes methods based on dynamic range stretch, global or local histogram equalization with or without histogram clip [Pizer, S.M., E.P. Amburn, J.D. Austin]

3. Edge sharpening is aimed at the underlining of boundaries and strokes. The most popular methods are based on the evaluation of Laplacian or gradient in different directions. The unsharp masking is a computer variant of a well known photographic technique.

4. Line refinement is of great use when disruptions in symbol's strokes or stroke merge are present. In many cases significant improvement could be achieved if mathematical morphology operations such as erosion, dilation, opening, closing, skeletonization and gradient evaluation are applied.

Image Segmentation

The goal of this operation is manifold [Shapiro, V., G. Gluhchev, V. Sgurev.].The first step is to extract text, i.e., to separate the text from the background. Provided the image is not too noisy and the background is uniform, a fixed threshold will be a fast and good solution. For such images the histogram is bimodal and the proper threshold corresponds to minimum between the two peaks. Unfortunately, in practice so nice images are rather exclusion than a rule. Very often images look like the one shown in Fig. 1a. In such cases the global threshold may either cause a significant loss of information or produce object–like artifacts, as shown in Fig. 2b. To avoid this, different locally adaptive methods have been developed, where specific threshold is evaluated taking into account the gray level distribution only in the predefined area, or in the neighborhood of every pixel. Fig. 1c demonstrates the effect of the application of the well–known Otsu method [Otsu N.].

The second step concerns the separation of rows. A very effective and cheap technique is based on the horizontal projections of the binarized image. The obtained shape has minimal values corresponding to the between–rows strips, while the peaks point out at the rows (Fig.3). To avoid false minimums due to ascenders or descenders, the shape has to be smoothed beforehand. The problem will aggravate if rows are not strictly

horizontal due to different reasons (Fig. 4). In that case techniques based on Hough transform [Lickforman–Sulem, L. and C. Faure] could be successfully applied. Thus, the angle of rotation could be evaluated and text rows might be rotated to become horizontal.

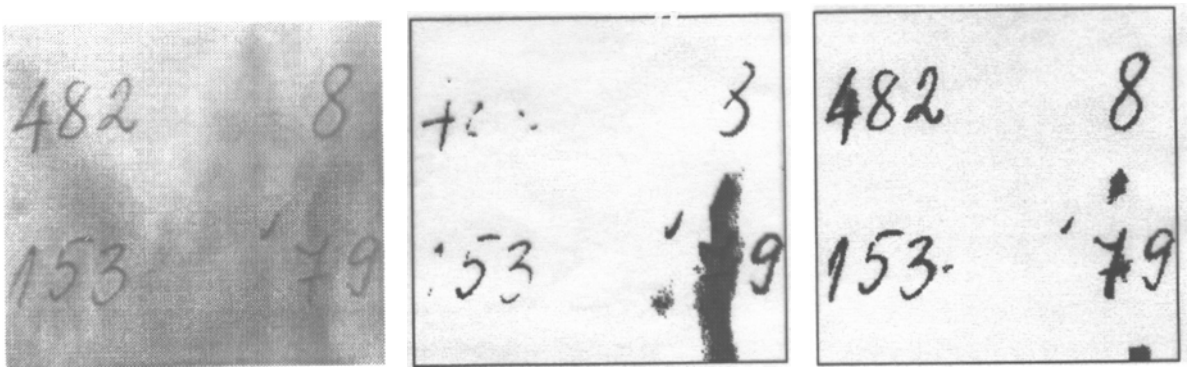

Fig. 2. a) Original noisy image b) Global threshold c) Locally adaptive threshold

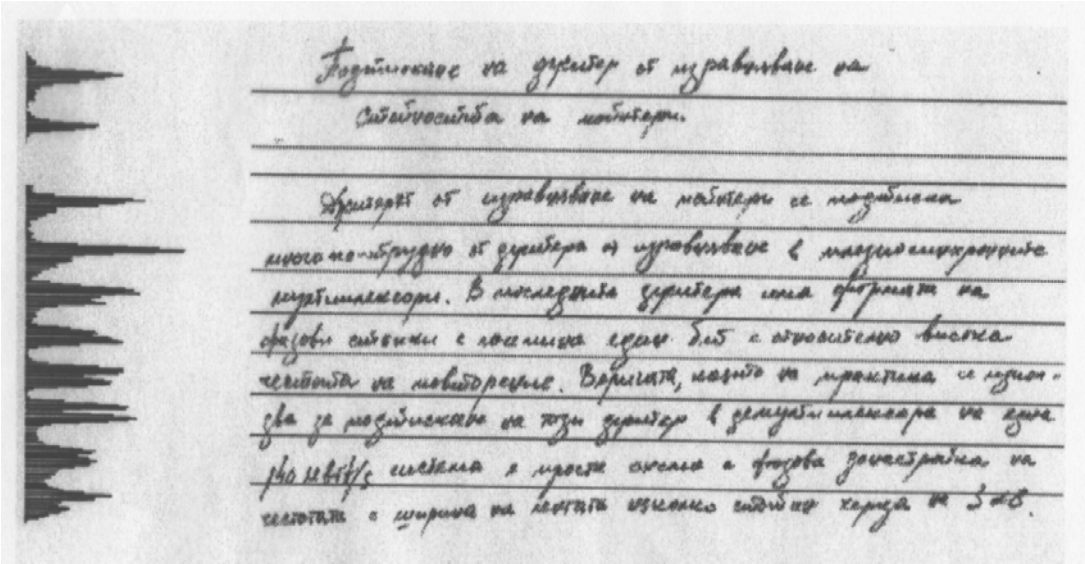

Fig. 3. Text lines separation

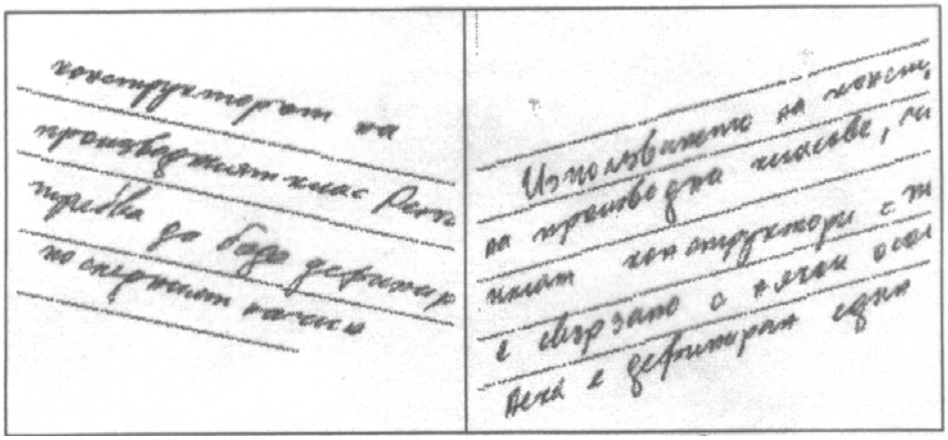

Fig. 4. Skewed text segmentation

The last segmentation stage concerns the separation of words, letters and strokes. While for the separation of words vertical projections of the rows could be successfully used, the problem of automatic segmentation of letters or strokes is quite complicated, especially for handwritten documents. Depending on the purpose of the investigation which may be text recognition, handwriting analysis or authentication, different processing is required. For example, in forensic investigations there is no need to recognize separate letters or words, but the goal is to establish document's or writer's authenticity. For this, specific features related to different letters, have to measured and compared. They may include distances between specific points, angles, curvature, as shown in Fig. 5.

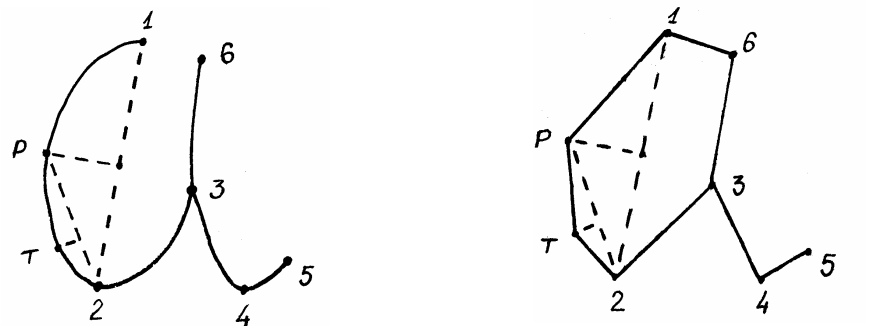

Fig. 5. Graphometric parameters of letter 'a'

Many approaches have been suggested for the recognition of words or letters. While the existing OCR packages perform very well on printed text, there is no reliable software for the recognition of scripts.

Current Projects of IIT

At present the text's processing and analysis research work in the IIT is carried out within following projects.

1. BioSecure – Network of Excellence, from the 6th European Framework Program
2. Biometric parameters based identification – Contract No I–1302/2003 with the Ministry of Education and Science (MES)
3. Fast access methods by content to multimedia databases – Contract No I – 1306/2003 with MES
4. Method and software for effective search by graphical content in large data–bases of images – Contract No ID6/2003 with the Ministry of Transport and Communications.

Acknowledgements

This work was supported by the Institute of Information Technologies and Ministry of Education and Sciences under contract No 1302/2003.

Bibliography

[Gluhchev G.] "Handwriting in Forensic Investigations", ICT&P, Varna, 2004 (in press)

[Lickforman–Sulem, L. and C. Faure]. "A Hough Based Algorithm for Extracting Text Lines and Handwritten Documents", ICDAR'95, Montreal, 1995, pp. 774–777

[Otsu N.] "A Thresholding Selection Method from Gray Level Histograms," IEEE Trans. Syst., Man, Cybern., vol.9, 1979, pp.62–66

[Pizer, S.M., E.P. Amburn, J.D. Austin] et al. "Adaptive hiastgram equalization and its variations",Comput. Vision, Graphics and Image Proc., 39, 1987, 355–368

[Pratt W. K.] Digital Image Processing, 2nd edn, John Wiley & Sons, 1991

[Shapiro, V., G. Gluhchev, V. Sgurev.] "Handwritten document image segmentation and analysis", Pattern recognition letters, North Holland, 1993, 14,pp. 71–78

Author Information

Georgi Gluhchev – Ph.D., Deputy Director of IIT, Acad. G. Bonchev Str., Bl. 2; Sofia 1113, Bulgaria.
e-mail: gluhchev@iinf.bas.bg