UDC: 519.711; 621.31.15

Data preprocessing for recognition of printed texts

Tea Todua

Georgian Technical University, 0175, Kostava 77, Tbilisi, Georgia.
Faculty of Informatics and Control Systems, Associate Professor.
tea_todua@gtu.ge

Abstract

Preparation processes of printed texts (smoothing and thinning) are presented. The smoothing method means converting background pixels into symbol pixels. Depth of smoothing depends on the vertical and horizontal dimensions of the raster. Converting rules are formulated on the basis of visual analysis of symbols. Proposed thinning algorithm carries out only one pass of raster. Elaborated algorithm provides symbol processing without distortions.

Keywords: *Data preprocessing, data preparation process, thinning, smoothing, Mini and Maxi-portraits.*

1. INTRODUCTION

In the early years of computer technology, it was realized that machine recognition of patterns was possible, and together with this arose the need of reducing to the minimum the amount of information necessary for the recognition of such patterns. It seems that the earliest experiments in data compression were conducted on character patterns in the 1950's. The thinned characters were used for recognition in [1], [2] and [3].

During these years, many algorithms for data compression by using of thinning have been devised and applied to a great variety of patterns for different purposes. In the biomedical field, this technique was found to be useful in the early 1960's, when a thinning algorithm was applied to count and size the constituent parts of white blood cells in order to identify abnormal cells. In other sectors, thinned images have found applications in the visual system of an automation, fingerprint classification, quantitative metallography, automatic visual analysis of industrial parts and etc. This wide range of applications shows that reducing of patterns and its representation as a thin-line is very useful. In addition, the reduction of an image to its essentials can eliminate some contour distortions while retaining significant topological and geometric properties. Naturally, for a thinning algorithm to be really effective, it should be ideally compress data, retain significant features of the pattern and eliminate local noise without introducing distortions of its own.

2. MINI AND MAXI PORTRAITS

As known, preparation is the pattern realizations transformation process, which purpose is to clearing up the realizations from different kind of obstructions. Owing to this realizations of the same pattern are obtained, which in a sense of reliability of their recognition are within admissible gradation area.

Elaborated method of preparation in the process of recognition is connected with using of standard descriptions, the so called Mini and Maxi Portraits [4]. Mini and Maxi Portraits are obtained by superposition procedures of binary pattern realizations of the learning set.

$$MAX_i = \bigcup_m \left\{x_{ni}^{\,mi}\right\};$$
$$MIN_i = \bigcap_m \left\{x_{ni}^{\,mi}\right\};\ \forall x_{ni}^{\,mi} = \{0,1\};\ \ n = \overline{1,N},\ i = \overline{1,I} \qquad (1)$$

MAX_i represents Maxi portrait of pattern A_i, MIN_i is Mini portrait for the same pattern, N is dimension of feature space, $I = Card\{A\}$, $m = Card\{X_i\}$.

If gradation of some pattern's realization is large, standard description of Maxi Portrait may embrace whole raster or his significant part, standard description for Mini Portrait may be equal to zero for the whole raster that makes difficult using of this method. To avoiding such situations there are necessary the following operations of preparation: smoothing and thinning.

3. SMOOTHING AND THINNING

a) Smoothing

By using of the smoothing procedure it is possible to reduce gradations of realizations of pattern on the edges of the raster. Elaborated algorithm depending on the raster size changes smoothing depth on the raster's edge. Besides of this, it is possible to avoid breaks caused by print and technological processes in the image structure (pic. 1a, 1b).

```
000000000111110000000000000
000000000111110000000000000
000000000111111100000000000
000000000111111100000000000
000000111111111111000000000
000000111111111111000000000
000000111111111111111000000
000000000111111111111000000
000000000111111111111110000
000000000001111111111110000
000000000001111111111111100
000000000000011111111111100
000000000000011111111111100
000000000000000111111111100
000000000000000111111111100
000000000000000000111111111
000000000000000000111111111
111111100000000000111111111
111111100000000000111111111
111111100000000000111111111
111111100000000000111111111
111111100000000000111111111
111111111000000000111111111
111111111000000000111111100
111111111000000111111111100
001111111000000111111111100
001111111111111111111111100
000011111111111111111110000
000011111111111111111110000
000000111111111111111000000
000000111111111111111000000000
```

```
000000000111111100000000000
000000000111111100000000000
000000000111111100000000000
000000000111111100000000000
000000111111111111000000000
000000111111111111000000000
000000111111111111111000000
000000000111111111111000000
000000000111111111111110000
000000000001111111111110000
000000000001111111111111111
000000000000011111111111111
000000000000011111111111111
000000000000000111111111111
000000000000000111111111111
000000000000000000111111111
000000000000000000111111111
111111100000000000111111111
111111100000000000111111111
111111100000000000111111111
111111100000000000111111111
111111100000000000111111111
111111111000000000111111111
111111111000000000111111111
111111111000000111111111111
111111111000000111111111111
111111111111111111111111111
111111111111111111111111111
111111111111111111111111111
111111111111111111111111111
```

a) Initial image b) Smoothed image

Pic. 1

b) Thinning

For the elaborated thinning algorithm the thickness parameter D is established in advance. This D parameter represents image thickness in pixels obtained by using of thinning. Proposed thinning algorithm carries out only one pass of raster.

The main problem was elaboration such procedure of thinning which provide elimination of interruptions in image (pic. 2, pic.3). As a result of breaks in image, symbol topology is destroyed: instead of image pixels appear background pixels.

Thinning method are relied on simultaneous considering of several rows (i, $i-1$and $i+1$) of raster. For avoiding breaks in image are used information about connections of rows. For this purpose are entered two parameters (α_1 and α_2) for which are performed two conditions: α_1=1, if $x_{ij}=1 \cap x_{i+1,j}=1$, in opposite case α_1=0. $\alpha_2=1$, if $x_{ij}=1 \cap x_{i-1,j}=1$, in opposite case $\alpha_2=0$.

```
11111111111111111111111111      11111111111111111111111111
11111111111111111111111111      11111111111111111111111111
11110000001111111111111111      11110000000111111111111111
11110000001111000000001111      11100000001110000000000111
11110000001111110000001111      11100000001111110000000111
11110000000011110000001111      11100000000011100000000111
11110000001111110000001111      11100000001111110000000111
11110000001111110000001111      11100000001111110000000111
11110000000011110000001111      11100000000011100000000111
11110000000011111000001111      11100000000011111000000111
11110000000011000000001111      11100000000011000000000111
11110000000000000000001111      11100000000000000000000111
11110000000000000000111111      11100000000000000000111111
11110000000000000001111000      11100000000000000000111000
11110000000000000001111000      11100000000000000000111000
11110000000000000000000000      11100000000000000000000000
11110000000000000000000000      11100000000000000000000000
11110000000000000000000000      11100000000000000000000000
11110000000000000000000000      11100000000000000000000000
11110000000000000000000000      11100000000000000000000000
11111111000000000000000000      11111111000000000000000000
00001111000000000000000000      00001111000000000000000000
00001111111111000000000000      00001111111111000000000000
00000000001111111110000000      00000000001111111110000000
00000000000000001111111111      00000000000000001111111111
00000000000000000000001111      00000000000000000000000111
00000000000000000000001111      00000000000000000000000111
00000000000000000000001111      00000000000000000000000111
00000000000000000000001111      00000000000000000000000111
00000000000000000000001111      00000000000000000000000111
00000000000000000000001111      00000000000000000000000111
00000000000000000000001111      00000000000000000000000111
```

a) Thinned image (D=4) b) Thinned image (D=3)

Pic.2

If conditions α_1=1 and α_2=1 are perform simultaneously or separately, then since this pixel, for which both conditions are performed, remain pixels which quantity is equal to D, other pixels will be erased.

If the image pixels are begin from the edge, then direction of thinning is from the centre of raster to the raster's edge.

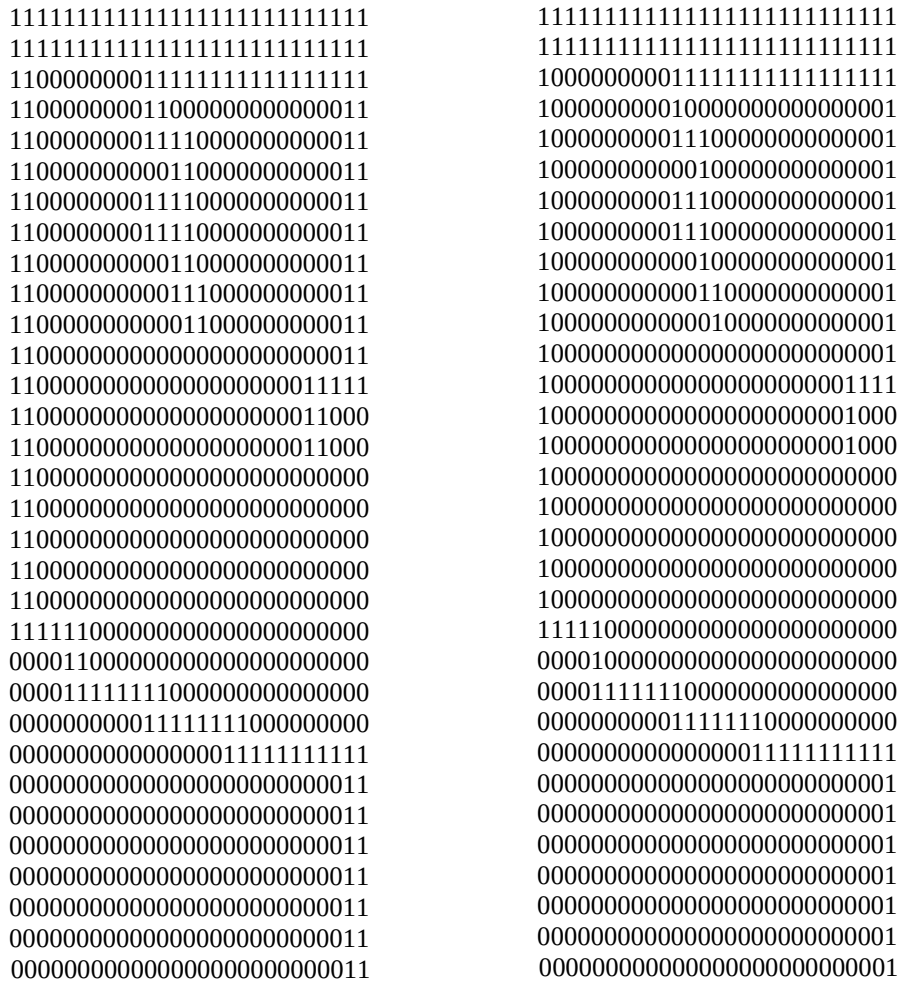
There are observed some limitations:

1. Long lines do not thinned.
2. If the image doesn't begin from the edge, then when choosing thinning direction there is considered how many successions of continuous units are in the given row.

There are considered the following situations:

1. There is the only succession of continuous units. In this case thinning takes place from right to left, if zero's quantity which are placed on the left side are more than right side placed zeros. If this condition is not performed, then thinning happens from left to right.

2. If succession of continuous units is equal to two in the given row, then the left group of the units becomes thinner from right to left, the right group from left to right.
3. If succession of continuous units are more than two in the given row, then the edge pixel's becomes thinner according to the rule indicated in paragraph 2. Middle located groups become thinner from right to left.



Pic.3

4. CONCLUSIONS

Smoothing procedure which are considered in this article, provides image's gradation reducing at the raster's edges. It is possible to avoid breaks caused by print and technological processes in the image structure. Elaborated thinning algorithm carries out only one pass of raster. Method of thinning relies on simultaneous considering of several rows of raster: the given row, the next row and the previous one. Both Thinning and Smoothing algorithms provide preliminary processing of image without distortions.

REFERENCES:

1. H. Sherman, “A quasitopological method for the recognition of line patterns”. in Proc. Int. Conf. on Inform. Processing (Paris, France), 1959, pp.232-238;
2. E.S. Deutsch. “Pre-processing for character recognition”, in Proc. IEEE NPL Conf. Patt. Recogn. (Teddington), 1968, pp.179-190;
3. T.M. Alcorn and C.W.Hoggar, “Pre-processing of data for character recognition”, Marconi Rev., vol. 32, pp. 61-81, 1969;
4. Verulava O. One logical decision-making procedure in pattern identification. Institute of control systems, Transactions. “Metsniereba”, Tbilisi, 1977;
5. Verulava O., Todua T., Gvichiani T., Zhvania T. About one Algorithm of stylization of Georgian Printed Symbols. Georgian Electronic Scientific Journal: “Computer Science and Telecommunications“. http://gesj.internet-academy.org.ge/gesj_articles/1292.pdf; #4(11), 2006;
6. Verulava O., Iremadze I., Todua T.. Preparation of Pattern Realizations by Neighbour Pixels Method. International Scientific Conference “Information Technologies 2008”. Technical University. Tbilisi, 2008.

Article received: 2008-11-09