Cursive character recognition – a character segmentation method using projection profile-based technique

Roberto J. Rodrigues

NCE- Núcleo de Computação Eletrônica/Universidade Federal do Rio de Janeiro,

Rio de Janeiro, Caixa Postal 2324, Ilha do Fundão, Brasil

Antonio Carlos Gay Thomé

NCE- Núcleo de Computação Eletrônica/Universidade Federal do Rio de Janeiro,

Rio de Janeiro, Caixa Postal 2324, Ilha do Fundão, Brasil

ABSTRACT

This paper reports the results of a study on a first sight decisi on tree algorithm for cursive script recognition based on the use of histogram as a projection profile technique. A postal code image data scanned is converted in a 2-dimension matrix representation to be used with a set of algorithms to provide full range segmentation. The results, based on this approach, are quite satisfactory for first stage classifier.

Keywords: Image Segmentation, Character Recognition.

1. INTRODUCTION

Document process applications can be found in almost all computer systems and now is become widespread. Applications like text edition, desktop publishing and graphics are often used for most organizations and home offices. This technology base has experimented a remarkable grown recently and besides the efforts in enhancements, all methodology still requires manual efforts to extract information, which means an exhausting task generally not fault-tolerant and time consuming.

The simulation of complex phenomena, mainly those related to nature, has been a big challenge for researchers. Vision process functions and visual patterns recognition, are fields of major interest for many of these researches.

Character recognition, as known as OCR (Optical Character Recognition) is an important subset within the pattern recognition area. OCR applications established some years ago the basis for the works within the research community in order to recognize and clarify pattern recognition and image processing analysis as an individual field of science.

The research of character recognition starts on the years of 1870's with the creation of the retina scanner. This device is an image transmission system with the use of a photocell mosaic. The sequential scanner created in 1890 was a real breakthrough in the development of the modern TV and optical reader devices. Character recognition itself had appeared as an aid system for blindness in the early 1900's. [The digital computer development at 1940 introduced the modern version of OCR developed, at first, only for a limited business data processing application.]

Recognition systems now face a paradox question: how to recognize without segmentation and, how to segment without recognition? The solution frequently used relies on the generation of many segmentation hypotheses, followed by tests applied over all possible combinations. Nevertheless, this strategy often leads to a combinatorial explosion with an obvious negative impact in the response time of the system (L.Duneau [5]). ICR (Intelligent Character Recognition) stands for a new proposal in this area of research and is gathering many adepts. This proposal implies the use of neural networks.

The first initiatives in the field of neural networks back towards the works of Norbert Wiener and John von Neuman in 40's [6]. The interest in this upcoming field decreased in the 60's, because of the related limitations presented by the existing network models (perceptron). At the end of 60's, mathematics theorem development and heuristic searching algorithm for chess game started the use of symbolic methods of IA. During 70's and the beginning of 80's, either symbolic method or expert systems formed the only methodology used for intelligent system implementation. However, by the 80's, it was noticed that the symbolic solutions were not so bright: the offered good performance under well defined conditions but they failed when the conditions were not very well known.

The best advantage of neural network comes from its ability to learn which means to self adjust upon the recognition of patterns based on a set of input data. The learning ability can be found with the presentation of a historical database. The network ability for learning and generalizing such relationship makes them more noise tolerant then other systems. The ability to represent non linear relationship makes them appropriate to a great number of applications, such industrial control system, computational vision and so on.

American post office service started using OCR systems by 1965. The system was not very reliable and presented operational problem very often. Brazilian Post Office ECT (Empresa de Correios e Telégrafos), deals with a volume of documents that has been experienced a substantial increase recently. Currently, it manages and delivers about 17 millions objects, including letters, orders, printed matters and telemessages. (http://www.correios.gov.br).

The goal of the research described in this paper is the construction an intelligent system for the automatic recognition of the Brazilian postal code (CEP – Código de Endereçamento Postal). The recognition process under investigation is divided into 5 main steps: image acquisition; image pre-processing; digit segmentation; neural network based recognition and finally, bar code generation.

2 Character recognition – problems and limitations

According to J. Mantas [2], script recognition can be classified as follows:

1. **Fixed-font character recognition**: It refers to the recognition of typewritten characters like pica, courier and so on.

2. **On-line recognition**: It is the method of hand-written character recognition where both the character image and the timing information of each trace are taken into account.

3. **Hand-written character recognition**: It refers to the recognition of typed hand-written characters.

4. **Script recognition**: It refers to those unrestricted handwritten characters that are cursive and may be connected.

The hardest and most complex of the classes is obviously the last one. There is no satisfactory technique for dealing with such cursive characters once shapes and individual cursive handwritten features cannot be limited into finite parameters.

The performance of an automatic recognition system depends on the quality of documents in its both forms original and digital. Many different approaches are used on the trial to compensate poor quality in the originals and in the captured images such as: contrast and noise level reduction. The problems related with quality and general picture handling are: [8]

1. *Noise* – unconnected line segments, pixels, curves etc.

2. *Distortion* – Local variations, rounded corners, improper extrusions and etc.

3. *Style variation* – Different shapes represents the same characters including type-like serif slants and so on.

4. *Translation* – It represents relative shift of the character. It can be entirely or by parts.

5. *Scale* – Relative size of the character

6. *Rotation* – Orientation changes.

7. *Texture* – Variations in the paper texture and the writing handler.

8. *Trace* – Variations in the thickness.

One of the most important and complex tasks in this process of individual character recognition is the segmentation. Nowadays, among some other known initiatives, the current one developed in CEDAR (Center of Excellence for Document Analysis and Recognition) at the University of Buffalo, represents a very interesting and exciting research.

The CEDAR major goal is to recognize the full address, including street, city, state and zip code. The study also includes foreign languages as Chinese, Japanese and Korean (http://www.cedar.buffalo.edu /).

Some other active groups of research in this area are:

IBM Pen Technology

(http://www.research.ibm.com/handwriting/)

NICI Handwriting Group, Nijmegen University, The Netherlands. (http://www.nici.kun.nl/)

Script and Pattern Recognition Group at the Nottingham Trent University

(http://152.71.57.102/Research/recog.html)

CEDAR, Document Recognition Group at SUNY Buffalo. (http://www.cedar.buffalo.edu/)

CENPARMI, Centre for Pattern Recognition and Machine Intelligence at Concordia, Montreal.

(http://www.cenparmi.concordia.ca/)

IDIAP, Artificial Intelligence group in Switzerland. (http://www.idiap.ch/)

DIMUND, University of Maryland.

(http://documents.cfar.umd.edu/)

OSCAR, Handwriting Recognition at Essex University (http://hcslx1.essex.ac.uk/)

Cursive character segmentation is in fact a very hard task and anyone who tries to attack the problem find himself faced to several unsolved problems, like slanted character segmentation, underlined and connected characters.

The literature presents any different techniques to face such problem, like contour analysis, incremental refinements [7], geometric and topological analysis, contour gradient [9], and so on. No one technique alone is able to solve all the previously cited problems and so, the major difficult is identify which we would give you robustness and generalization capabilities.

The investigation described in this paper has its major focus on the cursive version of the Brazilian zip code as used in mailing letters. Despite being a cursive problem, cursive writings of numerical patterns represent a slightly easier case than the generic cursive. By its nature, digits are (in most of the cases) written not in the connected form. The recognition process as conceived in this work, includes many stages as shown in the block diagram of figure 1 (attached), where segmentation is the phase of interest of this paper.

3 Segmentation process

The novel method, based on a decision tree construction and on the use of projection profile histograms, has been investigated and proposed.

Projection profile is a data structure used to store the number of non-background pixel when the image is projected over the normal X-Y axis (Eq. 1). Each cell of the projection vector is associated with the number of pixels above a predefined threshold (usually background color) (Eq. 2 and 3). An alternative projection histogram takes the average of the pixels intensity instead.

$$X,Y \rightarrow M(x,y) \tag{1}$$

$$X_n = \sum_{i=0}^{h} Y_i, n \in [0,v] \tag{2}$$

$$Y_n = \sum_{i=0}^{v} X, n \in [0,h] \tag{3}$$

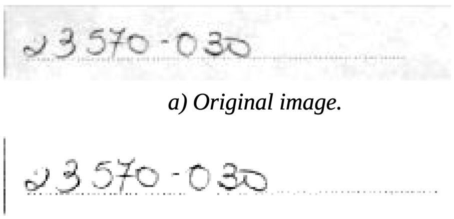
Where X and Y represent the horizontal and vertical axis, *h* represents the height of the picture (vertical size) for X or width (horizontal size) for Y and *v* represents the size of the picture.

The basic idea of the method consists on the construction of a tree with successive refinements, from the data of the histograms until a satisfactory performance is reached. Successive levels of the tree are allowed based on heuristic criteria. The algorithm includes three steps:

i. ***Image compensation***: This step is used in the trial to compensate the quality of the original image, enhancing certain details of the image as noise or contrast: It includes:

a) *Identification and background noise removal*: A low-resolution scanned image, not clean original or a colored envelope, certainly produces a poor result. For this type of image, a threshold factor will be necessary to remove the background color by filtering.

Figure 2 shows the background noise removed from a white paper zip code scanned with resolution of 200 dpi. As it can be seen, the removal process based on the projection histogram does not degenerate or even distort the original image, like it happens in some other filtering methods.



a) Original image.

b) Filtered image.

Figure 2: Background noise removal

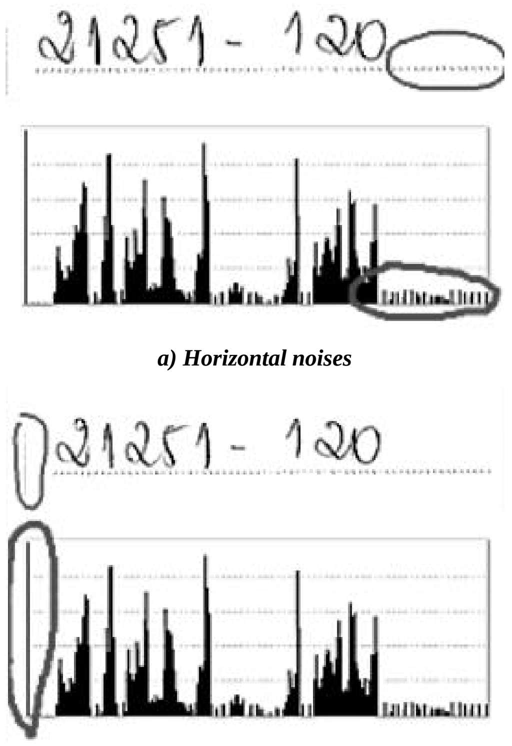
b) *Contrast enhancement*: Used to enhance bright images. (Figure 3)

a) Original image

b) Enhanced image

Figure 3: Contrast enhancement

c) *Removal of spurious pixels*: Can also be eliminated or reduced using the projection histograms (figures 4 and 5).



a) Horizontal noises

b) Vertical noises

Figure 4: Noise level histograms.

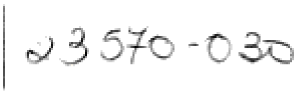
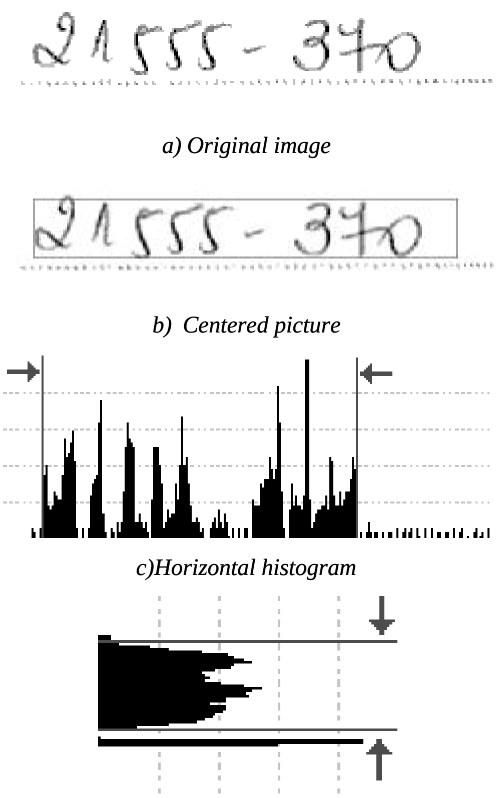


Figure 5: Dotted line removal.

d) *Cut*: The next step is to detach the central part where we can find the objects of interest, in this in case, the digits of the postal code (CEP). Again, based on the definition of some threshold bounds, the projection histograms provide the way to extract the image core from the rest (Figure 6).



a) Original image

b) Centered picture

c)Horizontal histogram

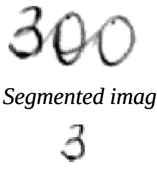
d) Vertical histogram

Figure 6: Image cut

Initial segmentation: In this step, a first segmentation is applied according to the information stored in the first level of the histogram structure. The segmentation is done based on the projected density of pixels and on an adaptive parameter called refinement rate.

Spurious pixels and lines can be removed using horizontal histogram data. Once the height of each digit influences the average of all heights, any element with a height smaller then a percentage of the average is a serious candidate to be removed. Another heuristic is used to identify segments with more than one possibly connected digit together. Such segments are selected for a new round with the next level of refinement.

ii. ***Refine:*** The second level of the decision tree is still based on the projection histogram, using a new refinement rate. Some weak connection can be broken in this level, as can be seen figure 7.



a) Segmented image

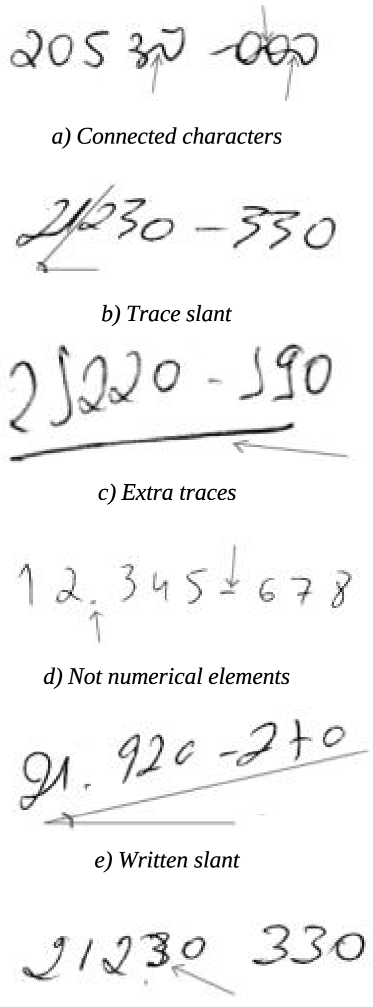
b) Digit 3 disconnected

c) Digit 0 disconnected

Figure 7: Refined segmentation.

The elements not successfully segmented, a new level of the decision tree is implemented using now a series of nom complex segmentation algorithms. The major point is that we can solve more than 80% of the general cases and 100% of the easy cases in a very fast and low computing way, using the projection histogram here described.

From the the segmentation viewpoint, the complicated cases are those the characters are connected to each other, or they are slanted, or they come with other spurious signs like bars, prints and underlines or even worse, when they are overwritten as shown in figure 8.



a) Connected characters

b) Trace slant

c) Extra traces

d) Not numerical elements

e) Written slant

f) Overwritten characters

Figure 8: Problems in the cursive written.

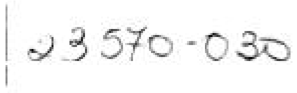
4 Obtained results

The experiment was carried out based on an image database specially generated within the community of the Federal University of Rio De Janeiro. Each five zip code data was collected from a different person belonging to different group based on their level of formal education, from first grade up to third.

The database was formed with 540 patterns of zip code. The images were scanned using 200 and 100 dpi, but only the 200 dpi ones were used because of the quality of the result. The color depth was defined as the standard RGB with 24 bits per pixel in gray scale. The final dimension of each scanned image (figure 5.2a) was around 500x120 pixels (200 dpi).

A dotted line was used to guide the writers what later, showed to introduce an extra level of difficulty to the problem.

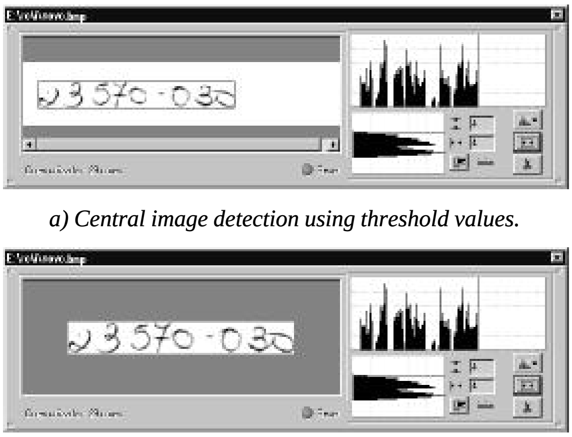
Once all images presented a high degree of noise and spurious signs, the first step applied was an image compensation. (Figure 11).



First result.

Figure 11: Noise removal.

After that, the image was cut. (Figure 12).



a) Central image detection using threshold values.

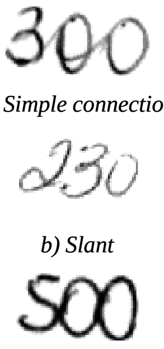
b) Central image detached.

Figure 12: Image cut.

From a total of 4320 digits, assuming 8 per zip code, the first level projection histogram algorithm extracted 3788 segments where 3286 were correct, 389 segments with multiple digits and 113 errors of segmentation (table 1 attached).

For this stage, it was used a value of 3 for the refinement rate and a value of $E0E0E0 (hexadecimal RGB) for the background color threshold.

Some of the multiple digits results include connected and slanted characters as shown in figure 13.



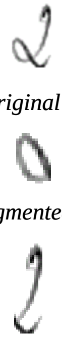
a) Simple connection

b) Slant

c) Cluster

Figure 13: Problems reported with multiple digits segments.

The most common type of error was observed to be an improper separation of a digit as if there was a connection (figure 14).



a) Original digit

b) Segmented part

c) Segmented part

Figure 14: Segmentation error.

Based on a simple heuristic, the algorithm separates the segmented objects into two sets: probably correct and probably multiple. The elements of the group were selected for the first refinement stage, with the rate of refinement set to 5. From 389 multiples digits, the algorithm was able to provide 320 correct segments, 48 multiple digit segments and 21 errors (table 2) The overall result, considering only these first two stages on the decision tree, achieved a performance of 83.47% of the 4320 expected digits.

5. Conclusions and future

The use of projection histograms for sure does not solve all segmentation problems, however it proved to be able to solve more than 80% of the cases in a very fast way. The method is easy to implement and for those were well formed zip codes it is able to provide 100% of accuracy in a short response time, what is an important issue considering commercial applications.

The use of special envelopes and guide to the writers would certainly increase the performance of the method.

The segmentation algorithm as proposed here has its basis on the compensation of the projection histograms and the use of an heuristic criterion to move on successive refinements. It becomes intuitive to conceive and design new levels to the decision tree with algorithms created to solve specific kind of situations.

A contour analysis algorithm is under current investigation to face slanted digits and also the use of a neural network for the heuristic part of the technique.

6 Bibliographical references

[1] D.G. Elliman, I. T. Lancaster, *A review of segmentation and contextual analysis techniques for text recognition,* Pattern Recognition, Vol. 23, No. 3/4, pp 337-346, 1990

[2] J. Mantas, *An Overview of Character Recognition Methodologies,* Pattern Recognition, Vol. 19, No. 6, pp 425-430, 1986

[3] W.H. Abdula, A.O.M Saleh, A. H. Morad, *A preprocessing algorithm for hand-written character recognition*, Pattern Recognition Letters 7 (1988) 13-18

[4] C. Y. Suen, M. Berthod, S Mori, *Automatic Recognition of Hand printed Characters,* Proceedings of the IEEE, Vol. 68, No. 4, April 1980

[5] L. Duneau, *Étude et réalisation d'un système adaptatif pour la reconnaissance en ligne de mots manuscrits,* Thèse de doctorat, Université Technologique de Compiègne, France, 1994

[6] J. Hertz, A. Krogh and R. Palmer, *An introduction to the Theory of Neural Computation*, ISBN 0-201-50395-6 and 0-201-51560-1 (1991).

[7] W. Verschueren, B. Schaeken, Y. R. de Cotret, A. Hermanne, *Structural Recognition of Handwritten Numerals*, CH2046-1/84/0000/0760$01.00@1984 IEEE

[8] C.Y. Suen, M. Berthold, S. Mori, *Automatic Recognition of Hand printed Characters – The State of The Art,* Proceedings of the IEEE, Vol. 68, No. 4, April 1980

[9] G, Srikantan, S. W. Lam, S, N, Srihari, *Gradient-Based Contour Encoding For Character Recognitio*n, Pattern Recognition, Vol. 29, No. 7, pp. 1147-1160, 1996

7. Figures

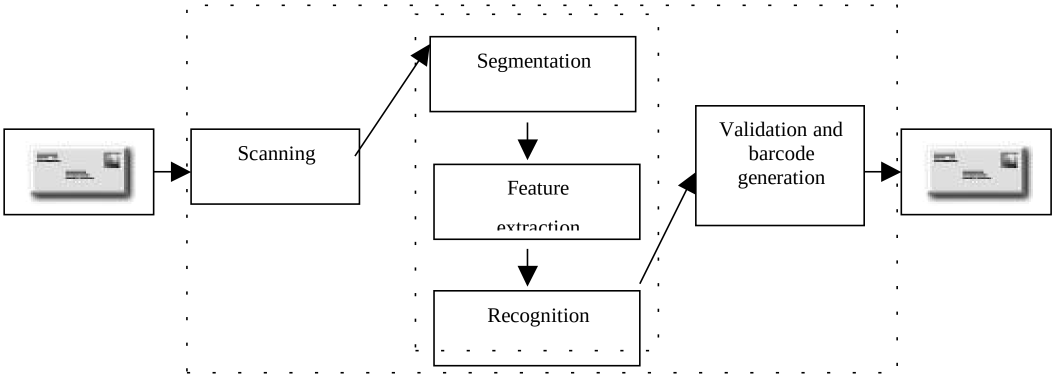


Figure 1: General diagram

Table 1: First segmentation.

	Quantity	% extracted	% expected
Straps	540	-	-
Expected digits	4320	-	-
Extracted segments	3788	-	-
Correct segments	3286	86,71	76,06
Multiple digits segments	389	10,26	9,00
Errors	113	3,00	2,60

Table 2: Second segmentation.

	Quantity	% extracted	% expected
Correct segments	320	86,26	-
Multiple digits segments	48	12,33	-
Errors	21	5,41	-

Table 3: Final result.

	Quantity	% extracted	% expected
Straps	540	-	-
Expected digits	4320	-	-
Correct segments	3606	95,20	83,47