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

ROBERTO J. RODRIGUES 1
ANTONIO CARLOS GAY THOMÉ 1
1NCE- Núcleo de Computação Eletrônica/UF RJ, Caixa Postal 2324, Ilha do Fundão, Rio de Janeiro, RJ, Brasil
cracky@nce.ufrj.br
thome@nce.ufrj.br

Abstract. This paper reports the results of a study on a first sight decision 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.

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 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.

\[ X, Y \rightarrow M(\pi, \gamma) \]  
\[ X_n = \sum_{i=0}^{h} Y_n, n \in [0, v] \]  
\[ Y_n = \sum_{i=0}^{v} X_n, n \in [0, h] \]

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.

![Figure 2: Background noise removal](image)

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

![Figure 3: Contrast enhancement](image)

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

![Figure 4: Noise level histograms.](image)

![Figure 5: Dotted line removal.](image)

d) **Cut:** The next step is to detach the central part where we can find the objects of interest, in this 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).

![Figure 6: Image cut](image)

**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 than 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.

![Figure 7: Refined segmentation.](image)

The elements not segmented successfully, in the succession of the refinements will be dealt with other segmentation methods in the sequence of the decision tree.

From the point of view of the segmentation, the more complicated cases are those that with connected characters, extra slant of the trace, ordinary traces, not-numerical elements, and overwritten characters as shown in figure 8.

![Figure 8: Problems in the cursive written.](image)

4 **Obtained results**

The experiment was carried through based on the information generated by image scanned in the academic community of the Federal University of Rio De Janeiro. The collection of data was made with a special form created, where each writer fills its personal information (age, grade, etc.) and writes 5 samples of the postal code (CEP-2 equal and 3 different ones). 540 straps of CEP had been used.

Initially, the pictures in the tests had been acquired from flatbed scanner at 200 and 100 dpi, however only 200 dpi images had been used. The color depth for the experiments in this paper was defined in standard RGB with 24 bits per pixel in gray scale. The final dimension of a CEP strap (figure 5.2a) is around 500x120 pixels (200 dpi).

540 straps had been used with CEPS written over a dotted line. This dotted line influenced negatively with the global performance. As the scanning process also picks-up the line; this infers a noisy element in all the pictures.

The scanned straps presented spots, eventual underlined traces and, since special color writing box form was not used. The first step of the sequence was spotted background removal, noise removal and the minimization of the dotted string effect. (Figure 11).

![First result.](image)

Next, the image was cut. (Figure 12).

![Figure 12: Image cut.](image)

With a sum of 4320 digits at all, assuming 8 by strap, the implemented algorithm extracted 3788 paths correctly, including...
segmented digits, the segmented subject with multiple digits and errors of segmentation itself, according to the statistics presented in table 1 (attached).

In this stage, the values of 3 for refinement had been used and of $E_{0E0E0}$ (hexadecimal RGB) for the threshold of background color.

The segmented subject with multiple digits includes scratch cluster resultant of slanted writing, connections, overlapping and overwriting. (Figure 13).

a) Simple connection

b) Slant

c) Cluster

Figure 13: Problems reported with multiple digits segments.

The most common errors include traces and points, which had passed by initial filtering, and digits badly segmented. The most common type of error observed was the cut of improper connection, as can be seen in figure 14.

a) Original digit

b) Segmented part

c) Segmented part

Figure 14: Segmentation error.

Based on a simple heuristic, the implemented algorithm separates the segmented objects amongst two sets: supposedly correct and supposedly multiple. These last ones are selected for the refinement stage. After the refinement with the modified initial values, 5 for the refinement degree, the program presented the results shown in table 2 (attached). With this result, we can build the table 3 presenting final result (attached).

5. Conclusions and future

The use of projection histograms for the process of character segmentation does not solve all presented problems, as trace slant for instance, however it does solve more than 70% of the cases with connected digits. Once the used algorithms has as base a very simple implementation, including the conventional filters, the process, in the whole, provide an excellent functional performance, taking the time necessary to carry through all the stages of the process.

The use of special envelopes, printed matters with colored delimiters, represent an element that certainly would raise consistently the performance of the method, because the delimiters easily would be removed from the scanned image by means of simple filtering, providing practically separated digits.

As the method of segmentation with projection histogram proposed in this paper it have all decision heuristic formed on the basis of the refined density from the picture, it becomes sufficiently intuitive the conception of new algorithms as leaves to be inserted in the tree.

Thus, the future steps include the improvement of the method, mainly the algorithms of contour analysis to enumerate the connected characters and the development of new algorithms to solve the problems reported here. Another way to enhance the performance of this method, as well as of the algorithms of the succession, is to rely on the bed of a neural network for the validation of the resultant segmented characters. The improvement of the performance would be given in the reduction of significant stages in the sequence of the process from the network response.

6 Bibliographical references


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


7. Figures

Figure 1: General diagram

Table 1: First segmentation.

<table>
<thead>
<tr>
<th>Quantity</th>
<th>% extracted</th>
<th>% expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Straps</td>
<td>540</td>
<td>-</td>
</tr>
<tr>
<td>Expected digits</td>
<td>4320</td>
<td>-</td>
</tr>
<tr>
<td>Extracted segments</td>
<td>3788</td>
<td>-</td>
</tr>
<tr>
<td>Correct segments</td>
<td>3286</td>
<td>86.71</td>
</tr>
<tr>
<td>Multiple digits segments</td>
<td>389</td>
<td>10.26</td>
</tr>
<tr>
<td>Errors</td>
<td>113</td>
<td>3.00</td>
</tr>
</tbody>
</table>

Table 2: Second segmentation.

<table>
<thead>
<tr>
<th>Quantity</th>
<th>% extracted</th>
<th>% expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct segments</td>
<td>320</td>
<td>86.26</td>
</tr>
<tr>
<td>Multiple digits segments</td>
<td>48</td>
<td>12.33</td>
</tr>
<tr>
<td>Errors</td>
<td>21</td>
<td>5.41</td>
</tr>
</tbody>
</table>

Table 3: Final result.

<table>
<thead>
<tr>
<th>Quantity</th>
<th>% extracted</th>
<th>% expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Straps</td>
<td>540</td>
<td>-</td>
</tr>
<tr>
<td>Expected digits</td>
<td>4320</td>
<td>-</td>
</tr>
<tr>
<td>Correct segments</td>
<td>3606</td>
<td>95.20</td>
</tr>
</tbody>
</table>