A New Cost Function for Typewritten Digits Segmentation

C. Rodríguez, J. Muguerza, M. Navarro, A. Zárate, J.I. Martín, J.M. Pérez
Computer Architecture and Technology Department
The Basque Country University (UPV/EHU)
Aptdo. 649, 20080, Donostia, Spain
E-mail: acprolac@si.ehu.es

Abstract

This work presents a solution to the problem of the segmentation of digits in forms characterized by its low quality, as well as the existence of breaks and touching digits. We propose a new function of segmentation that adds to two traditional techniques (vertical projections and Tsujimoto metric) information of background of the digit. Unlike other techniques reported in the literature, ours obtains a near-optimum number of break points in fields containing broken, blurred and touching characters, leading to high accuracy in the global OCR system. The accuracy obtained in the segmentation of the forms fields is of 99,74% on a sample of 11,283 fields of 144 forms of low quality, which provides a final accuracy to the automatic recognition process of 99,42% of digits correctly classified.

1 Introduction

An Optical Character Recognition (OCR) system is composed by several phases [1]. The segmentation phase, which is the goal of this paper, plays a determinant role in the global accuracy of the OCR system. The complexity of the segmentation is present in the great variety of sources and the characteristics of the image, such as the bad quality of the printing, the existence of problems in the digitalization process, the bad conservation of documents to process, etc. Furthermore, most of OCR systems work with binary images obtained after transforming an image into grey scale using a certain threshold. This process habitually generates breaks in the characters and unions between contiguous characters (touching). Considering these characteristics, the degree of segmentation complexity can be established from the nature of the text to be processed: (a) well-formed characters, uniformly or proportionally spaced, (b) broken and touching characters, (c) typed characters in cursive similar to handwritten, and (d) handwritten characters.

In this paper, we are dealing with the segmentation of numeric fields in forms presenting significant blurring, breaking and touching. The segmentation mechanism we propose has been implemented as part of an OCR system that is operating in diverse Spanish institutions (like the Bank of Spain and "Colegio Nacional de Registradores"), as well as in diverse private companies. More details of the complete OCR system can be consulted in [4] and [5].

The paper is arranged as follows: in Section 2 we present a new cost function for segmentation. Our proposal is evaluated in Section 3, and, finally, the Section 4 presents the conclusions.

2 A new segmentation function

There is not a consensus about a technique of general application [2]. For that reason, combined techniques are usually applied in some systems. In addition, does not exist a universal segmentation criteria [3]. In most cases the segmentation accuracy is based on the recognition results obtained with that segmentation. Thus, the selection of a classifier will influence the results.

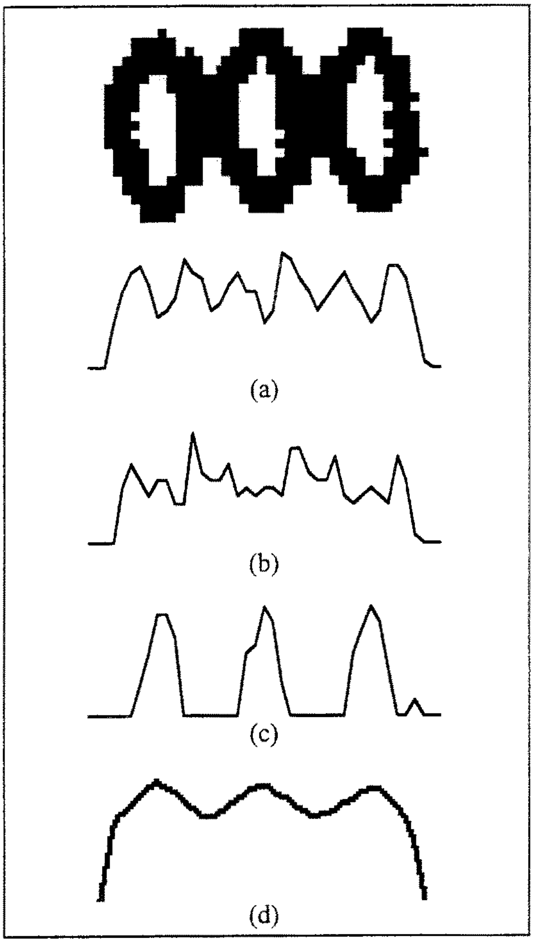

Fig. 1. The three components of the cost function: (a) vertical projection, f_1, (b) Tsujimoto metric, f_2, and (c) the new f_3 component. (d) The cost function f_c after smoothing.

Considering the methods that make the segmentation independent of the recognition, the most extended method consists of making the *vertical projection* or histogram of black pixels, that lies in calculate the sum of the existing number of black pixels for each column of the field. The valleys of this function represent points of cut in the field that is going to be segmented, as it can be see in Figure 1a. This technique has some troubles with touching characters, as we can see in the same Figure.

In order to solve these drawbacks, Tsujimoto [6] proposed an alternative metric based on the connectivity among the pixels of the bitmap (Figure 1b). Tsujimoto metric surpasses vertical projection in touching character segmentation but it fails when addressing the segmentation of blurred and broken digits like the ones in Figure 2. To manage these situations, we consider *background information* of the digit image to introduce in the cost function, f_c, a new component (Figure 1c), f_3, along with the calculated vertical projection, f_1, and Tsujimoto metric, $f2$. The new function component tries to correct the not wished minima of the function (those that correspond with the inner part of the digits) produced by $f1$ and $f2$, introducing the pixels that comprises the interior of the number. Figures 1d and 2b shows the three-component f_c after smoothing.

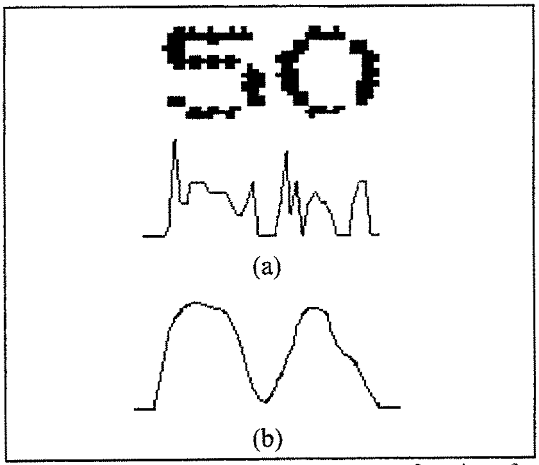

Fig. 2. Tsujimoto metric (a) and the 3-component cost function after smoothing (b).

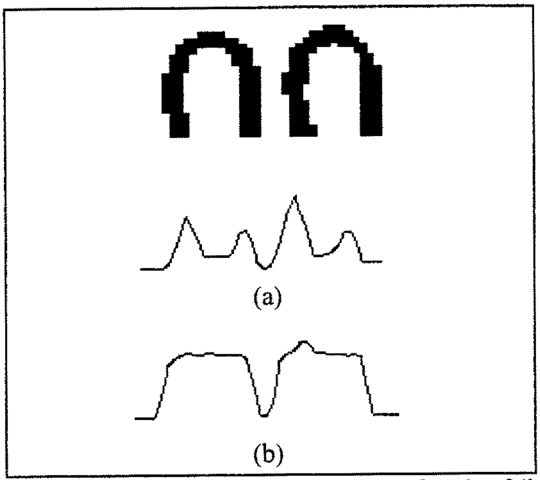

Fig. 3. A case of breaking where the 3-component cost function fails (a).The new 4-component cost function (b).

Nevertheless, some kinds of breaks in digits, like the one in Figure 3, are not well solved by the above three-component cost function (Figure 3a). To cope with these situations, very frequent in form recognition, we introduce a fourth component, f_4, in the cost function. The component f_4 increases f_c when f_3 is zero and f_1 is not zero. The introduction of component f_4 brings a new, four-component, cost function that is the basis of our proposal. Figure 3b illustrates the result of the new f_C for the same example field.

The equations corresponding to the components of the cost function are the following ones, where x_{ij} is the value of a pixel of the image (0/1):

Component	Definition
1st	$f_1(x_{ij}) = 1 \Leftrightarrow x_{ij} = 1$
2nd	$f_2(x_{ij}) = 1 \Leftrightarrow x_{ij-1} = x_{ij} = x_{ij+1} = 1$
3rd	$f_3(x_{ij}) = 1 \Leftrightarrow x_{ij} = 0 \wedge (\exists x_{kj} \neq 0 \wedge \exists x_{rj} \neq 0)$ with $k < i,\ r > i$
4th	$f_4(x_{ij}) = 1 \Leftrightarrow x_{ij} = 0 \wedge \exists x_{kj} \neq 0$ with $k < i$

Finally, the cost function (for each column j) is a pondered sum of the four components:

$$fc_j = \alpha_1 \sum_i f_1(x_{ij}) + \alpha_2 \sum_i f_2(x_{ij}) + \alpha_3 \sum_i f_3(x_{ij}) + \alpha_4 \sum_i f_4(x_{ij})$$

where α_1=2, α_2=3, α_3=1.5 and α_4=2. The values of α parameters have been obtained from the experimentation made with our data.

In order to eliminate not wished points of cut, after calculating the cost function for each column, a global smoothing is made. In this process, the equations (1) and (2) have been applied. On the basis of the experimentation made, the number of times that there are to apply each one of the equations has been determined: the equation (1) is applied only one time, and the equation (2) is applied 30 times. At the end, the smoothing that produces a smaller variance of size of the box candidates was chosen.

$$fc_j = \frac{fc_{j-1} + fc_{j+1}}{2} \qquad \forall j \; if\, (fc_{j-1} > fc_j < fc_{j+1}) \qquad (1)$$

$$fc_j = \frac{fc_{j-1} + fc_j + fc_{j+1}}{3} \qquad \forall j \qquad (2)$$

3 Evaluation of the cost function

A sample including forms supplied by the Bank of Spain has been used for evaluation purposes. These forms, filled up by diverse Spanish companies, include a great variation of sources and styles of printing. In addition, the quality of the printing is very variable. In average, each form includes 500 digits, with about 80 numeric fields. In the selected companies the characteristics of blurring, breaks and unions between digits are dominant. For this reason, it can affirm that the obtained results can be considered slanted negatively.

From the previous estimation, the test made with a sample of 144 forms, 11,283 numeric fields and 53,017 digits, shows that the 99,74% of the fields have been correctly segmented. Only 50 digits have presented problems in the segmentation: 19 pairs of digits have been merged and 12 broken ones. Another important result is that, with respect to the segmentation, 88,2% of forms do not present any error.

In Figure 4, some additional examples to show the robustness of the method are illustrated.

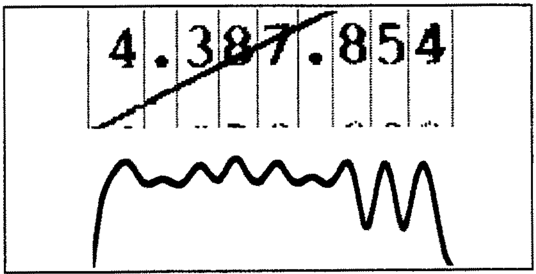

Fig. 4. An example showing the robustness of the cost function.

4 Conclusions

This article has presented a new mechanism of segmentation that combines the metric of Tsujimoto, the method of the vertical projection and two new components based on the information of background of the field. The presented mechanism has been developed as part of an OCR system oriented towards recognising blurred and broken typewritten digits in forms.

The presented function of cost generates an almost optimal number of points cut candidates in the fields to segment, that are characterized by their low quality (blurred fields and with cuts). The experiments made show that only a 0,26% of the fields of the selected sample present errors of segmentation. On the basis of this good accuracy of the segmentation process, the complete system of OCR developed (operative, for example, in the Bank of Spain) obtains a recognition accuracy of 99,42% [5].

Furthermore, the cost function is being applied to the segmentation of handwritten digits and alphanumeric characters, with promises results.

5 Acknowledgements

This work was supported in part by the EU PASO Project No. 002.4/93 *Form-Less,* whose partners are ADHOC Synectic Systems S.A., DOC6 S.A., Registro Mercantil de Tarragona, Barcelona Activa S.A., Computer Architecture and Technology Department (UPV/EHU), and Pentha Magnetics S.A.; and by Gipuzkoako Foru Aldundia and University of the Basque Country.

6 References

[1] **R.G. Casey, E. Lecolinet:** A Survey of Methods and Strategies in Character Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 7, pp. 690-706, 1996.

[2] **A.C. Downton, R.W.S. Tregidgo, E. Kabir:** Recognition and Verification of Handwritten and Hand-printed British Postal Addresses. *Character & Handwriting Recognition*, Ed: P.S.P. Wang, pp. 265-291, World Scientific series in Computer Science Vol. 30, 1991.

[3] **Y. Lu:** Machine Printed Character Segmentation - An overview. *Pattern Recognition*, Vol. 28, No. 1, pp. 67-80, 1995.

[4] **J. Muguerza:** *Una Solución al Reconocimiento Automático de Dígitos Imprecisos en Formularios*. Doctoral Thesis, Basque Country University, Spain, January 1996.

[5] **C. Rodríguez, J. Muguerza, M. Navarro, A. Zárate, J.I. Martín, J.M. Pérez:** A Two-Stage Classifier for Broken and Blurred Digits in Forms. Accepted for presentation in the 2nd *International Workshop on Statistical Techniques in Pattern Recognition*, Sydney, Australia, 1998.

[6] **S. Tsujimoto, H. Asada:** Resolving Ambiguity in Segmenting Touching Characters. *The First International Conference on Document Analysis and Recognition*, pp. 701-709, 1991.