Analysis of a Neural-net-based Algorithm for the Segmentation of difficult-to-read Handwritten Letters.

PILAR GÓMEZ-GIL and JORGE NAVARRETE-GARCÍA
Department of Computer Science
Universidad de las Américas, Puebla
Santa Catarina Mártir, Cholula Puebla. 72820
MÉXICO
pgomez@mail.udlap.mx

Abstract: - Today, the automatic recognition of printed documents is a common task due to the success of OCR´s. However, the interpretation of old handwritten documents is still an open problem. Segmentation of words in characters is one of the most difficult tasks in this automatic recognition process. In this paper we present the results obtained by an algorithm proposed by [1], when applied to words extracted from twenty five different telegrams written by General Porfirio Díaz, former president of México, during his government at the beginning of the 20th century. This algorithm combines two different heuristic algorithms to generate Possible Segmentation Points in a word, which are then reselected by an Artificial Neural Network (ANN).

Key-Words: - Handwritten recognition, word segmentation, Artificial Neural Networks

1 Introduction

The process of recognition of an isolated manuscript word is composed by several steps: image extraction, preprocessing (elimination of noise and writing variants), word segmentation and character recognition [2]. Segmentation is the operation that seeks to decompose a word image in a sequence of several images containing isolated characters [5]. When the images to divide present too many ornaments, damages or missing characters due to the nature of the original documents, segmentation becomes a very difficult task [3].

The work presented here uses a technique originally defined by Blumestein and Verma [1], that combines heuristic methods with intelligent methods. We applied this technique to the segmentation of difficult-to-read handwriting words taken from telegrams and letters written by Porfirio Díaz during his ruling in Mexico at the beginning of the 20th century. These documents are very important for the History of México, because they reflect the political and economical situation lived at that time, and they contain important information about the revolution movement.

2 Generation of words

The purpose of this part of our research was to find a useful algorithm to segment Diaz's words, therefore the rest of the processes for automatic recognition are not described here.

To obtain the words, 25 telegrams were scanned, and the images cleaned from noise and printed lines using commercial software, manually cutting a set of black and white isolated word images. Figure 1 shows some examples of these words. In part (a) of the figure it is clear that reading turns difficult because it is hard to find out the beginning and end of each character. It can be noticed in (b) the different ways in that character “a” can be written, depending upon the position it takes in the word, even though it was written by the same person. Word (c) shows that in some cases characters ”n”, “l”, “e”, “u” and “ii” can be written all in the same way.

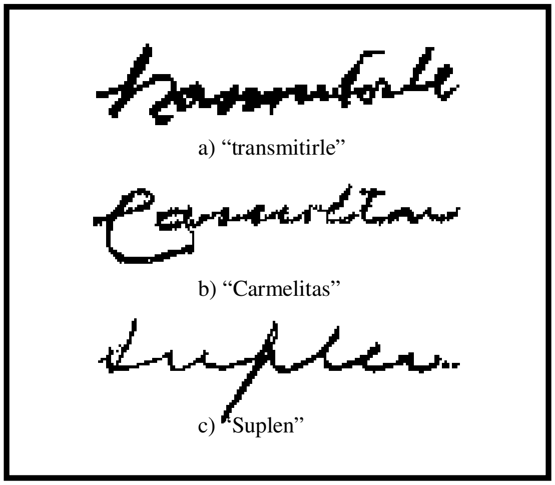
3 Segmentation Algorithm

The analyzed algorithm was proposed by Blumenstein and Verma [1]. In our opinion its value comes from the combination of heuristic and intelligent methods. First two heuristic segmentation algorithms are implemented to

Published at WSEAS Transactions on Systems. Issue 4, Vol. 3. June 2004. ISSN: 1109-2777.pp. 1426 - 1429. Not for massive distribution.

generate possible segmentation points (further referenced as PSP). After that, these points are analyzed by a Neural Net trained with back propagation that will decide if they are valid. The implemented heuristic algorithms are White Holes Segmentation [5] and Vertical Pixel Density Algorithm [4]. The first algorithm detects white pixel areas rounded by black runs, in order to find caves or circumferences corresponding to letters as *a,b,c,d,e,g,h,n,o,p,q*. This method forces segmentation points before and after the white area. The second algorithm builds a black pixel density histogram for each word column so that valleys in the histogram indicate the presence of a ligature between letters. The use of more than one segmentation method produces an over segmentation, that is, the PSP exceed the number of characters contained in a word. To eliminate over segmentation and find the correct segmentation points, a trained neural net is used.

The network is trained using a set of classified correct and incorrect segmentation points. Each segmentation point is represented by a set of windows extracted form a binary matrix derived from a black and white word image. A window represents the black pixel density found in a 5*5 pixels area. Figure 2 shows an example of two windows.


a) “transmitirle”

b) “Carmelitas”

c) “Suplen”

Figure 1. Examples of Porfirio Diaz’s handwriting

A friendly software application, called HAWOST (Handwritten Word Segmentation Tool), was built to train the network and segment the words. This tool is divided in two main panels: the first is dedicated to create, train and test the net, and the second panel takes care of the segmentation process and the use of the current ANN for classification. Several networks can be tested and stored. Using this tool, the researcher can visually appreciate the performance of each heuristic algorithm and its success when it is combined with others. After segmentation the use of the trained ANN becomes available, and it allows the researcher to estimate how segmentation is improved. Figure 3 shows a snapshot of this tool.

4 Results

Three major experiments were carried out. The first uses very hard-to-read words, like words with short ligatures, overlapping, and bad quality, they require an expert interpretation. The second uses an ANN trained with "easy" words, like words with prominent ligatures easy to read for common people. The third uses a training file with easy and hard words combined.

All experiments use the same network topology (270-300-200-100-1), a learning rate of 0.12. 200 epochs were used to train the networks.

The complete segmentation process was tested with 20 difficult words. Results are shown in table 1.

As expected, training the network with difficult words improves de number of correct segmentation points when difficult data is tested. The low level of over segmentation obtained in the three cases demonstrates the success of the hybrid technique in this type of writing.

5 Conclusion

From this work, it is concluded that when hard-to-read words need to be segmented, a good strategy consists in the use of more than one segmentation algorithm, obtaining the best of each method.

However the use of many algorithms produces over segmentation. To eliminate this, a neural network can be used, training it to distinguish correct from incorrect segmentation points.

The results presented here can be improved adding more heuristic segmentation algorithms and training ANN’s with a larger number of patterns.

References:

[1] M. Blumenstein, B. Verma: A New Segmentation Algorithm for Handwritten Word Recognition. International Joint Conference on Neural Networks, Washington D.C. , (1999) 878-882

[2] Pilar Gomez Gil, Sergio Linares Perez, Carlos

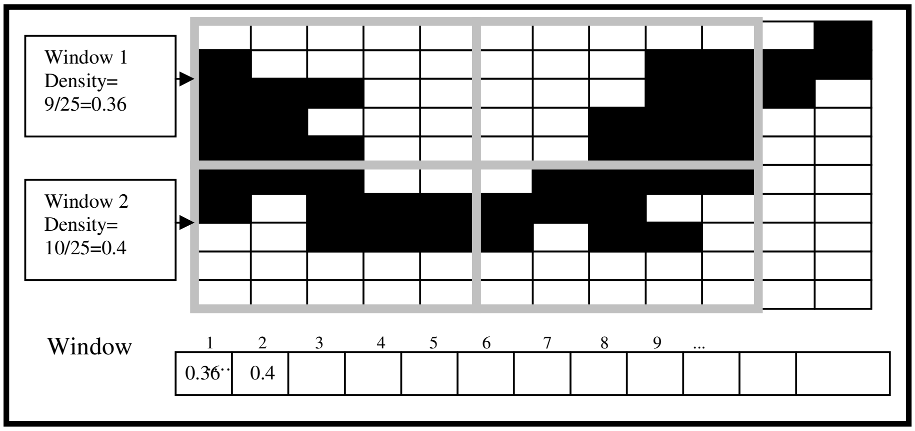
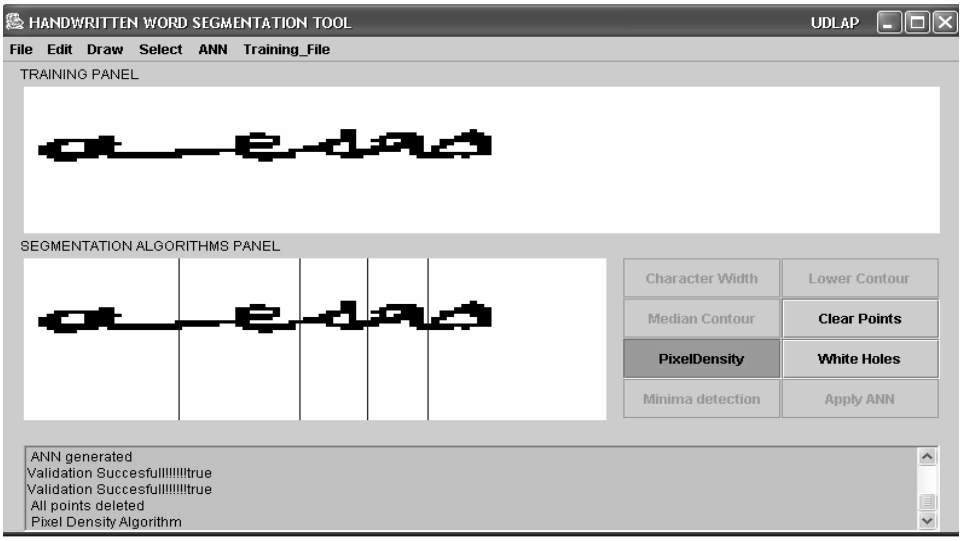


Figure 2. Generation of Windows



```
ANN generated
Validation Succesfull!!!!!!true
Validation Succesfull!!!!!!true
All points deleted
Pixel Density Algorithm
```

Figure 3. A view of HAWOST

Table 1. Hybrid segmentation results on difficult-to-read handwriting

Training input	Ideal number of segmentation points in the test set (*D*)	Correct segmentation points found by HAWOST (*C*)	Incorrect segmentation points found by HOWOST (*I*)	Over-segmentation (*C*+*I*-*D*)	Over-segmentation rate (*C*+*I*-*D*)/*D*
Hard-to-read words	82	67	22	7	8%
Easy-to-read words	82	43	23	-16	-19%
Mixed words	82	55	18	-9	10%

Spínola, Manuel Ramírez Cortes: On the Automatic Digital Storage of Historical Documents: Recognition of Handwritten Telegrams of Don Porfirio Díaz. Knowledge-Based Intelligent Information Engineering Systems and Allied Technologies, (2001) 753-757.

[3] Pilar Gómez-Gil, Cristian Castañeda, Sergio Linares, Carlos Spínola and Manuel Ramírez: Reconocimiento de Letra Manuscrita para la Creación Automática de Bases de Datos Digitales. Conferencia Iberoamericana en Sistemas, Cibernética e Informática, (2002) 115-118

[4] Kussul Mikhailovich, E. and Kasaktina, L.M : Neural Network System for continuous handwritten Words Recognition. Book of Summaries of International Joint Conference on Neural Networks. Washington, D.C., (1999) 22

[5] Nicchiotti G., Scagliola, C., Rimassa. S. : A Simple and Effective Cursive Word Segmentation Method. Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition, Amsterdam, (2000) 499-504

Published at WSEAS Transactions on Systems. Issue 4, Vol. 3. June 2004. ISSN: 1109-2777.pp. 1426 - 1429. Not for massive distribution.