The Role of Neural Networks in the Interpretation of Antique Handwritten Documents

Pilar Gómez-Gil, Guillermo De los Santos-Torres, Jorge Navarrete-García, Manuel Ramírez-Cortés

Department of Computer Science, (2) Department of Electrical Engineering. Universidad de las Américas, Puebla. MEXICO

Abstract. The need for accessing information through the web and other kind of distributed media makes it mandatory to convert almost every kind of document to a digital representation. However, there are many documents that were created long time ago and currently, in the best cases, only scanned images of them are available, when a digital transcription of their content is needed. For such reason, libraries across the world are looking for automatic OCR systems able to transcript that kind of documents. In this chapter we describe how Artificial Neural Networks can be useful in the design of an Optical Character Recognizer able to transcript handwritten and printed old documents. The properties of Neural Networks allow this OCR to have the ability to adapt to the styles of handwritten or antique fonts. Advances with two prototype parts of such OCR are presented.

1 The Problem of Antique Handwritten

Currently, web distribution of old documents is limited to a scanned image of the document because most of the commercial Optical Character Recognizers (OCR) do not obtain good recognition rates with old handwritten documents or with documents using old styles of fonts.

The recognition of old handwritten and printed documents is a challenge in pattern recognition, due to special characteristics that this recognition problem presents. Figure 1 shows an example of an old telegram, written by Gral. Porfirio Díaz, president of Mexico at the beginning of XX Century. Even for a non-expert person, who does not have some previous knowledge of this kind of writing, is very difficult to interpret the content of this document.

P. Gómez-Gil et al.: *The Role of Neural Networks in the Interpretation of Antique Handwritten Documents*, StudFuzz **208**, 269–281 (2007)
www.springerlink.com
© Springer-Verlag Berlin Heidelberg 2007

Digital processing of old documents faces, among others, the following conditions:

- Old documents have been damaged with the pass of time. In most cases they present spots, color of paper have changed, or their texture is deteriorated.
- Digitalization process requires special cares to protect the documents. The production of a digital image that will feed the OCR is, by itself, a delicate process. It requires a special kind of scanner, which would not touch the document.
- The recognition process of old documents is off-line. There is no information about the dynamics of the writing or the pressure used by the writer.

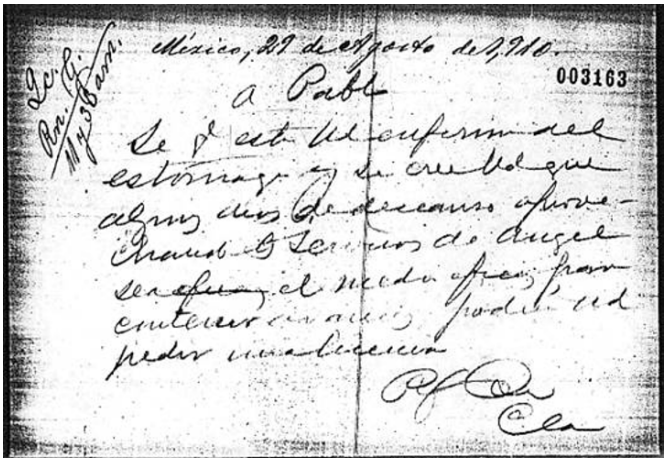

Fig. 1. An example of a telegram written by Porfirio Díaz at the beginning of XX century [1]

Added to these conditions, there are also special complications during the recognition of old handwriting. Some of them are [2]:

- Old styles of handwriting have a lot of ornaments.
- Fonts are not uniform. For example, same character may look different in different places of a word, in different words or in different documents. Notice that this situation is presented in any kind of handwriting, and is much stronger if documents came from different writers.
- The shape and style of writing may be different even for the same person depending on environmental factors, mood, type of pens, age, etc.

- Character segmentation requires extra procedures, besides the common ones as identification of valleys and hills, due to the styles of different letters.
- In some patterns it is noticed that different classes of characters are very similar in shape.

Figure 2 shows some examples of handwritten words written by the same writer at different moments and documents. Notice that some letters have different shapes depending on their positions in the word, and when presented in different words. Some letters may be confused with a connection and some letters may be "embedded", looking two of them as one character. Therefore, in terms of a pattern recognition problem we have that:

- There are no evident prototypes to define each class
- The variance among members of the same class is greater than expected values
- Common similarity metrics, as Euclidian distance, are sometimes useless because it may be greater for patterns belonging to same class than for patterns belonging to different classes.

2 An OCR for Antique Handwritten Documents

The research group of Neural Networks and Pattern Recognition at Universidad de las Américas, Puebla, is currently working with the construction of an OCR able to recognize antique handwritten and printed documents. This OCR will be useful to our library, which posses a huge amount of such historical documents [3].

We propose the construction of an adaptive OCR, called *Priscus* (latin word meaning "antique") that have the following components (see figure 3):

– Digitization. Creation of a color or gray level image of the document to be recognized.
– Pre-processing. Cleaning of image, noise reduction and black and white conversion of the image.
– Segmentation of words. Given a binary map, this process obtains the words that are presented in the image.
– Segmentation training. Adaptive system that learns to identify segmentation points in a word, based on the handwriting or font presented to the OCR.

- Character segmentation. This process obtains possible segments that may contain characters, based on the knowledge obtained from the segmentation training.
- Recognizer training. Adaptive system that learns to identify characters from segments obtained from the binary image of the document.
- Recognition of characters. It receives segments of words, extract features of them and decide the most likely characters.
- Identification of words. Based on possible words obtained by the recognizer and a dictionary, this process decides the most likely words.
- Correction of style. Based on the identified most likely words and grammar rules, this process creates well formed sentences, obtained a transcription of the document.

At this point, we have focused our research in the segmentation and character recognition components, using artificial neural networks.

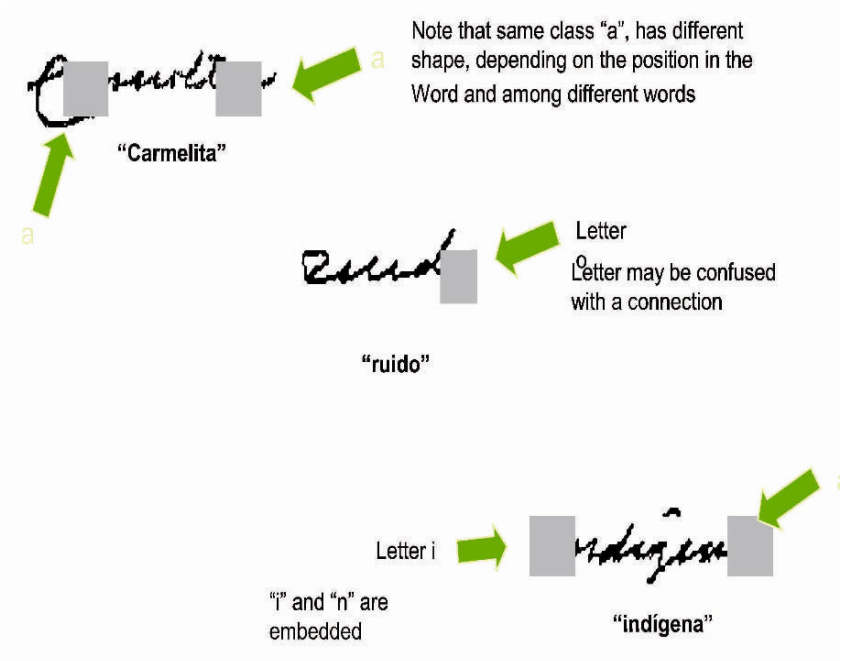

Fig. 2. Examples of old handwritten words [2]

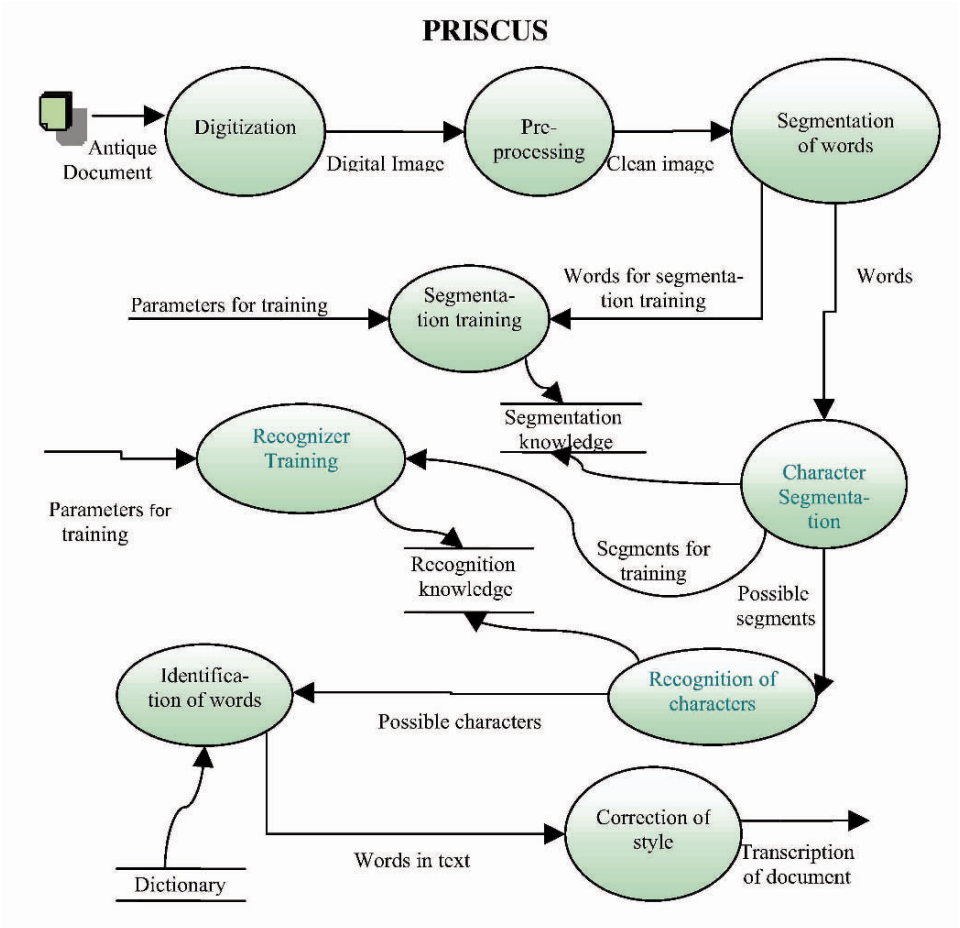

Fig. 3. The proposed OCR for transcription of old documents

3 Why Neural Networks?

Artificial Neural Networks (ANN) are mathematical models inspired in biological systems, able to learn the behavior of a system getting their knowledge from data. For our purposes we use ANN simulated in digital computers, but they are available also as hardware components. There are many types of ANN, and most of them present three important characteristics: abstraction, generalization and learning [4]. ANN are useful for prob-

lems were the functions describing a system are not explicit, but there is enough data that likely can be used to obtain some numerical representation of the knowledge ruling the behavior of the system.

It is evident that the definition of explicit rules for the segmentation and recognition process in handwriting is a very complex task, therefore the use of adaptive systems that learned from examples for such processes are useful. In the other hand, it is known that some models of neural networks are able to obtain better generalization rates than others adaptive systems. This generalization ability is an advantage in the case of handwriting recognition, given the spread presented in the patterns of the clusters of classes presented in this case. In the other hand, having trainable systems, segmentation and recognition can be tailored to the style of the writing found in specific applications or for fonts found in specific periods of the history.

Up to date, we have been working with the development of prototypes of two parts of the OCR: a system based on a back-propagation neural network [5] for segmentation of words and a character recognizer based on a SOFM neural network [6]. Both sub-systems will be explained in the next sections.

4 Test Case: Telegrams Written by Gral. Porfirio Diaz

The library of the Universidad de las Américas Puebla, contains a rich collection of old documents. Among them, there are about 70,000 telegrams written during a historical Mexican epoch known as "*Porfiriato*." The library has the goal of making them available across the web. Up today, around 2,000 of them have been digitized and their contents transcribed by experts and they are available for consulting [1].

In order to test our segmentation and recognition systems, 25 of such telegrams were scanned, and their images were manually cleaned from noise and printed lines using commercial software, and manually a set of black and white isolated word images were cutting.

5 Segmentation Using a Back Propagation ANN

We build a segmentation application, called HOWOST, based on 3 components: the "white hole algorithm" proposed by Nicchiotti et al. [7], a vertical density algorithm proposed by Kussul & Kasaktina [8] and a Back propagation neural net trained to reduce the over segmentation generated by both algorithms.

The white hole 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 BP neural network is trained to learn which of the segmentation points generated by both algorithms are right and which are wrong. The supervised classification to train the network is given by a human expert that marked the correct and incorrect segmentation points in some examples automatically generated for the algorithms. After the network is trained with examples of the handwritten or font documents, the system may be used on-line to segment words. Figure 4 shows the interface of the segmentation subsystem.

Different configurations of networks with different data sets have been tested, but three major experiments were carried out to test the performance of the system. The first experiment uses very hard-to-read words, like words with short ligatures, overlapping, and bad quality, requiring an expert for their interpretation. The second experiment uses an ANN trained with "easy" words, like words with prominent ligatures easy to read for common people. The third experiment uses a training file with easy and hard words combined. All experiments use the same network topology (270-300-200-100-1), and a learning rate of 0.12. Training was stopped after 200 epochs. Table 1 shows the results of these experiments. 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.

In general, using a set of 898 mixed patterns, we got 83% of accuracy in segmentation combining the two algorithms with the neural network. When tested independently, the white hole algorithm obtained 47% of success in the best case, and the vertical density algorithm obtained 53% .

6 Self-Organized Maps for Character Recognition

For the recognizer we chose a neural network able to create topological maps trained with a non supervised algorithm [10]. We decided to use topological maps because, given the special characteristics of these problems, it was mandatory to have several prototypes of each class, and to relate them in a way that similar prototypes were near in a way that their

relative relation were shown. The decision of use non supervised algorithm was based on the idea that the learning showed by humans when reading old handwriting is no supervised.

Inspired in the organization by maps of human brain T. Kohonen developed the self organizing feature mapping algorithm (SOFM) [6]. The goal of SOFM algorithm is to store a set of input patterns $\mathbf{x} \in X$ by finding a set of prototypes $\{\mathbf{w}j \mid j = 1, 2\ldots N\}$ that represent the best feature map $\mathbf{\Phi}$, following some topological fashion. The map is formed by the weights connection $\mathbf{w}_j$ of a one or two-dimensional lattice of neurons, where the neurons are also related each other in a competitive way.

This learning process is stochastic and off-line; that is, two possible stages are distinguished for the net: learning and evaluation. It is important to notice that the success of map forming is highly dependent on the learning parameters and the neighborhood function defined in the model. The map is defined by the weights connecting the output neurons to the input neurons.

Following is a description of the SOFM algorithm as applied in the construction of the recognizer [11].

1. Initialize the weights with random values:

$$\mathbf{w}_j(0) = \text{random}() \quad , j = 1..N \text{ (number of neurons)} \tag{1}$$

2. Chose randomly a pattern $\mathbf{x}(t)$ from the training set X at iteration t.
3. For each neuron i in the map feature map $\mathbf{\Phi}$ calculate the similarity among its corresponding weight set $\mathbf{w}_i$ and $\mathbf{x}$. The Euclidian distance may be used:

$$d^2(\mathbf{w}_i, \mathbf{x}) = \sum_{k=1}^{n} (w_{ik} - x_k)^2 \quad \text{i} = 1..N \tag{2}$$

4. Find a wining neuron i^* which is the one with maximum similarity (minimum distance).

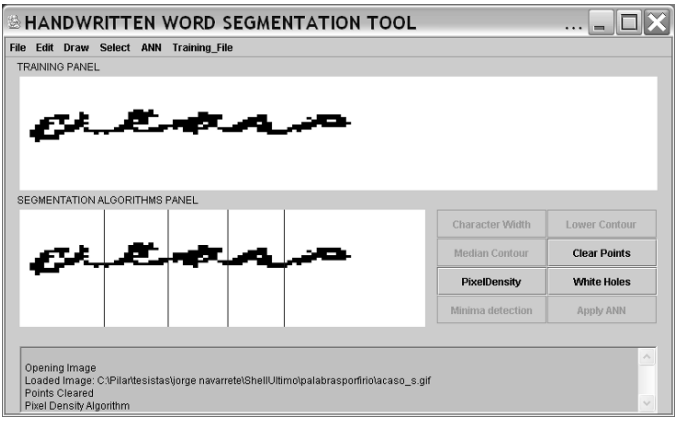

Fig. 4. Our segmentation software [9]

Table 1. Results obtained by the segmentation subsystem

Training input	Ideal number of segmentation points in the test set	Correct segmentation points found by HAWOST	Incorrect segmentation points found by HOWOST
Hard-to-read words	82	67	22
Easy-to-read words	82	43	23
Mixed words	82	55	18

5. Update the weights of winning neuron i^* and their neighbors as:

$$\mathbf{w}_j(t+1) = \mathbf{w}_j(t) + \alpha(t)(\mathbf{x}(t) - \mathbf{w}_i(t)) \quad \text{for j} \in \Lambda_{i^*}(t) \tag{3}$$

6. Where $\Lambda_{i^*}(t)$ corresponds to a neighborhood function centered on the winning neuron. For this problem, we choose a neighborhood distance

of 0 neurons. $\alpha(t)$ is a learning rate function depending on time. We choose: $\alpha(t) = 1/t$.

7. Go to step 2 until no more changes in the feature map are observed or a maximum number of iterations is reached.

Several experiments have been made with a different number of classes in order to analyze and understand the behavior of this network. We started with 3 classes up to 21 classes. Unfortunately, at the moment of this work, we did not have enough data to test the whole alphabet with 27 classes (the whole alphabet). The results of SOFM were compared with a recognizer based on a "nearest neighbor algorithm" using a "k-means" algorithm to get the prototypes required by nearest neighbor as described at [12]. Table 2 shows the results obtained by both the SOFM network and the nearest neighbor classifier. Notice that in all cases the SOFM network gets better results that the nearest neighbor algorithm.

Figure 5 shows some topological maps generated by the SOFM using patterns of the five vowels. Notice that the maps result as expected. Similar prototypes are generated near each other. Figure 6 shows the topological maps for 21 classes generated by the network.

Table 2. Results of recognizer with different number of classes [10]

Number of classes	Number of training patterns	Type of Recognizer	Recognition rate on Training set
3	13	Nearest neighbor	84%
		SOFM (3x3)	92%
5	56	Nearest neighbor	58%
		SOFM (5x1)	58%
		SOFM (5x2)	71%
		SOFM (5x5)	73%
21	86	Nearest neighbor	6%
		SOFM (5x12)	63%
		SOFM (2x30)	70%

7 Conclusions and Future Work

We presented the overall issues associated to the construction of a OCR able to process antique handwritten and printed documents. The special characteristics associated to this problem were discussed, as well as the role of artificial neural networks in the implementation of useful segmentation and recognition systems. The advances obtained in these two subsystems were also presented.

It is clear that there is still a lot of work to be done, because each component of this recognizer is by itself a complete system. At this moment we are working with the integration of segmentation and recognition subsystems, as well with the identification of a systematic way to look for the best SOFM topology. It can be noticed that we did not present results from the application of the recognizer or the segmentation system in old printed documents; we are also working with this part.

Acknowledgments

Authors would like to thank Mr. Alberto García-García, for the advisements given during the development of this work, as well as for providing us with the images of the telegrams.

Fig. 5. Topological maps generated for vowel using different topologies [10].

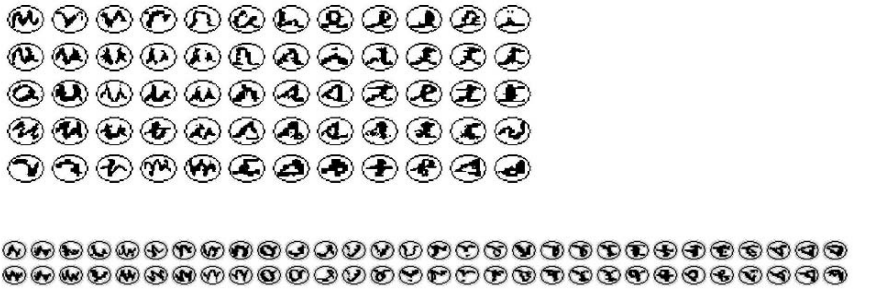

Fig. 6. Topological maps generated for 21 classes using 2 different topologies [10].

References

1. Universidad de las Américas, Puebla. Digitalización, Codificación y el Acceso Vía Internet de los Telegramas del ex presidente de México Porfirio Díaz. In: Colecciones Digitales Biblioteca (2002) http://biblio.udlap.mx/telegramas
2. Gomez-Gil, P.; Navarrete-García, J.: Analysis of a Neural-net based Algorithm for the Segmentation of Difficult-to-read Handwritten Letters." In: WSEAS Transactions on Systems. Issue 4, Vol. 3 (2004) 1426 – 1429
3. García-García, A.: Digitalización y Divulgación Digital de Acervos Antiguos. In: Servicios Digitales. Bibliotecas de la Universidad de las Américas Puebla. http://ict.udlap.mx/projects/cudi/buap/ (2004)
4. Haykin S.: Neural Networks: a Comprehensive Foundation. Macmillan College Publishing Company. New York. (1994)
5. Rumelhart, D.E. G. E. Hinton and R.J. Williams.: Learning Internal Representation by error propagation. In: Parallel Distributed Processing: Explorations in the Microstructure of Cognition D.E. Rumelhart and J.L. McClelland, eds. Vol. 1, Chapter 8. Cambridge, MA: MIT Press. (1986)
6. Kohonen, T.: Self-Organized formation of topologically correct feature maps. Biological Cybernetics, 43, (1982) 59-69.
7. 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.
8. 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.
9. Navarrete-García, J.: Mejora en el algoritmo de segmentación para el reconocimiento de caracteres de telegramas escritos por el Gral. Porfirio Díaz.

Tesis para obtener el grado de Licenciatura. Departamento de Ingeniería en Sistemas Computacionales. Universidad de las Américas, Puebla. (2002).
10. De-los-Santos-Torres, G.: Reconocedor de Caracteres Manuscritos. Master thesis. Departamento de Ingeniería en Sistemas Computacionales. Universidad de las Américas, Puebla. (2003).
11. Gómez-Gil, Pilar, De los Santos-Torres, M., Ramírez-Cortés, Manuel: Feature Maps for Non-supervised Classification of Low-uniform Patterns of Handwritten Letters. Progress in Pattern Recognition, Image Analysis and Applications, Lecture Notes in Computer Science Vol. 3287 (2004) 203-207.
12. Tao, J.T. and Gonzalez, R.C. Pattern Recognition Principles. Addison-Wesley (1974)