A Hierarchical Classifier for Multifont Digits

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

In this paper, the automatic recognition of broken and blurred, multifont typewritten digits in forms will be addressed. The classification, which is based on the utilization of a global feature, is divided in two phases: first, a minimum distance method (1-NN) is applied to provide a global classification of the patterns in a form; second, the patterns in the form previously classified are used to validate, or reject and reclassify them, on the basis of the mean distance to the predefined classes. In this way, a classification accuracy rate of 99.42% has been achieved.

1 Introduction

The importance of the development of the computers is known in the data processing, mainly including problems that given to their excessive volume of data or its necessity of calculation was not able to manage until the moment. Thus, the importance of the capacity of acquisition and " understanding " on the part of the computers is a fundamental aspect and in continuous study [9,14,20]. In this sense, a system of Optical Character Recognition (OCR) has as objective the conversion of independent form of characters written to codes that represent them and that are intelligible to the computer. Typical examples of applications of a OCR system are the automatic introduction of data in commercial surroundings [16], as well as the reading and classification of ZIP and address in mailing [12,19].

Nowadays there are systems of OCR able to recognize as typed texts as written by hand, although the low quality of documents to recognize and the great variability of fonts cause that the problem is still without solving of global way [6,7]. This is due to the deficient quality of the images of the digits to recognize (broken and blurred characters), or deformations originated by a bad conservation of the original document, or due to a bad digitalization or later segmentation of the same one. In this article both problems (multifont and broken and/or blurred digits) in the scope of the recognition of typed digits are considered.

The experimental sample has been provided by the Bank of Spain in the frame of the European project Form-Less and come from forms of an average of 500 digits filled up annually by 250,000 different Spanish companies. For the development of the prototype, whose commercial version is operative in the Bank of Spain and others Spanish companies, a sample has been taken from 63,700 digits pertaining to 161 different sources. These sources have been selected by disparity of fonts and by present a low quality of printing (noise, dirt, bad digitalization, etc.). For that reason, the presented results are skewed negatively with respect to the global accuracy of the system in real surroundings.

A peculiar aspect of this application, which we create general in the use of forms, is the property monofont of the forms of a company, reason why in the phase of classification we raised a solution that taking advantage of this property obtains results that maintain the level of error below the 0,7%. This percentage is considered

maximum error for this application, because manual intervention will be excessively onerous with greater error rates, making the application non-practical.

The structure of the paper is the following. Section 2 reviews the different classifiers that we have considered in order to make the experimentation explained in this paper. Section 3 presents the proposed two-phase classifier and the experimental results, and Section 4 is devoted to showing the conclusions.

2 Classifiers

Without trying to make an analysis exhaustive of all the existing methods of classification, we have made a series of experiments that have allowed us to discriminate between diverse classifiers for our concrete application. Before happening to expose in the following section the obtained results, we will present these classifiers briefly.

2.1 Parametric Classifiers

Within this group include those classifiers that are based on discriminant functions that they assume that the function density of the population is known and some parameters must be estimated. In particular, we have worked with a classifier that we will denominate *multivariate linear classifier*, where the density function is a normal one, and the parameters to consider are the average and the variance. The equation of the discriminant function, g(), comes given by the following expression:

$$g(X_j) = -(X_j - m^{(i)})^t * S^{-1} * (X_j - m^{(i)})$$

where *Xj* is the pattern to classify (in our case, the vector of features of the digit), *m(i)* is the average sample of class *i* and *S* is the variance sample.

2.2 Non-Parametric Classifiers

These methods do not estimate any knowledge of the density function. We have selected the k-Nearest Neighbor method [4]. In this method, there exist several metrics to apply in order to determine the ranges between the patterns: absolute, Euclidean or Mahalanobis. Also, in the case of $k>1$, as far as the definition of the weight, exists diverse variants [5] to assign *j*-th closest neighbor, w_j and the following criterion based on the relation between the distances, d, of the k patterns nearest pattern *j* has been chosen:

$$w_j = \begin{cases} \dfrac{d_k - d_j}{d_k - d_1}, & d_k \neq d_1 \\ 1, & d_k = d_1 \end{cases}$$

2.3 Artificial Neural Networks

Within this group, we have worked with two paradigms: *feedforward* networks and the *Restricted Coulomb Energy* model. From the first group, a multilayer perceptron has been taken [11,18] with the backpropagation algorithm [17]. As far as the second paradigm, the commercial version developed by Nestor Inc. has been used [2]. It is a model oriented to the recognition of patterns that is based on the generation of prototypes with a certain radius of influence that is modified during the training.

For this type of classifiers several tests have been made varying the different parameters (learning factors, number of hidden units, radios of influence of the prototypes, etc). The following section shows the best obtained results.

3 The Hierarchical Classifier

As mentioned in the Introduction, our classifier operates in two phases. A general, previously trained, *multifont* classifier carries out a first classification of the digit images; then a second, *specialized*, one-form oriented classifier, validates the former classification and tries to reclassify the rejected patterns.

3.1 Phase One: Multifont Classifier

In order to select the feature to be extracted from the digit image, experiments have been carried out to evaluate the features reported in the literature [14,15]. In general, the features can be classified in two great groups according to their nature: *global*, they are based on the association of patterns examining the image of a global form, and *structural*, treat the image analyzing topological aspects (for example, holes, contour, skeleton, etc.). Within the first type, we have worked with the *zoning* feature [1,3,10], that consists of dividing the box of the digit by zones and extract a vector of features on the basis of the black percentage of pixels that has each zone. With regard to the structural features, we have used the *skeleton*.

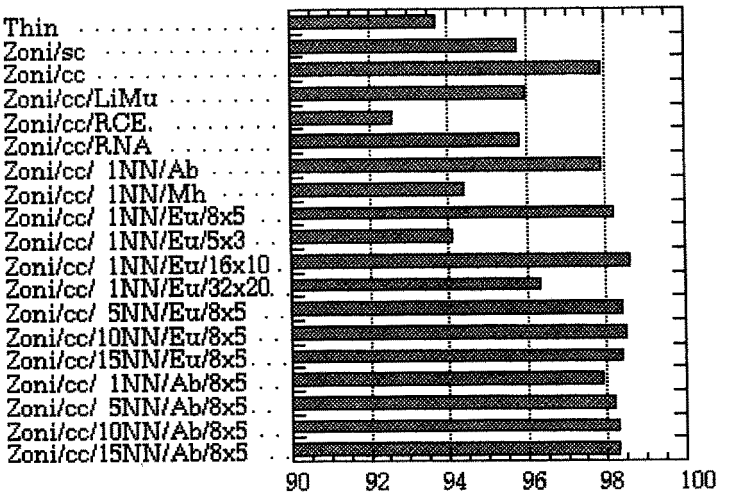

Fig. 1. Accuracy of the tested classifiers.

The selection between zoning and the skeleton has been made on the basis of a 1-NN classifier due its good accuracy for other applications of recognition of characters [8,13]. Figure 1 shows the accuracy obtained with both types of features. As we can see, the best result of the skeleton (Thin in Figure 1) is below the accuracy of zoning. This is due to the greater accuracy obtained by zoning in comparison with the skeleton for blurred and broken digits, in addition to its rapidity of extraction.

At this point, an additional preprocessing was introduced. As we observed that more than 25% of the digits in the sample are broken at the lower side, the height of the digits was statistically estimated on the basis of the median of the digits in the form. From this estimation, the portion of the digit that is really relevant is established and the remaining rows were ignored in classification (Figure 2). This technique improves the ratio of digits correctly classified. In our experiments, this ratio increases from 95.78% (Zoni/sc in Figure 1) to 97.89% (Zoni/cc in Figure 1).

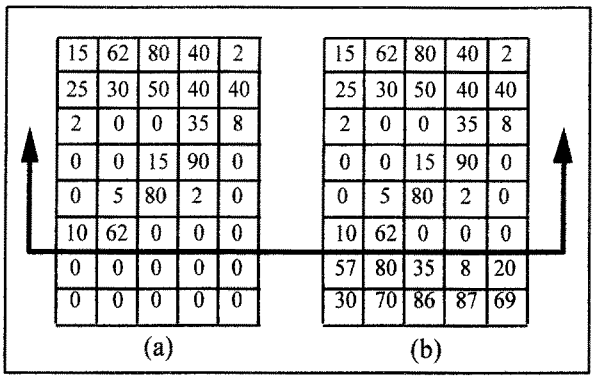

Fig. 2. A broken digit (a) and its the reference pattern (b). The two lower rows will not be considered for the classification.

Using the zoning feature and the technique previously explained, we have analyzed the classifiers studied in Section 2 varying the corresponding parameters. In particular, the metric of distance between patterns (Ab: Absolute, Mh: Mahalanobis, Eu: Euclidean), the degree of neighbor of the k-NN (1NN, 5NN, 10NN and 15NN) and the dimensionality of the vector of features (8x5, 5x3, 16x10 and 32x20). As show in Figure 1, the best results have been obtained by k-NN classifier, in concrete using the Euclidean metric, with k=1 and a 16x10 grid. For other values of k the results have been similar, being more expensive by the maintenance of a ordered list of k elements. With regard to dimensionality, although a 16x10 grid obtained a slight better accuracy rate, a grid with 8 rows and 5 columns of dimension was found to be the most suitable due to the much lower dimensionality of its feature vector (40 against 160).

The percentage of success that is obtained in this phase is of 98,16, with a 1,84% percentage of errors or substitutions. It should be remarked that the classification was not fully satisfactory, as we established the error limit in 0.7% for the reference application (as explained in the Introduction). For that reason, it was necessary to improve the classification by means of a second phase.

3.2 Phase two: Specialized Classifier

Any classifier oriented to multifont forms presents the problem of bad classification by similarity between different classes from different fonts. This interference has been reflected in the study of the classifiers analyzed in the previous section. This problem is become serious in applications like which it concerns to us, in that when having of the order of 250,000 different forms, the amount of fonts is innumerable.

Obviously the disadvantage of the interference is diminished when the problem is reduced to classify patterns monofont. In the application that we are treating, we see that practically each one of the forms to recognize presents the same source. For that reason, we raised a specialized classifier that taking advantage of this circumstance reduces, in this phase of recognition, the problem multifont to monofont.

This classifier consists of two subphases. First, starting off of the patterns previously classified with a certainty (by means of the classifier multifont of the previous phase), a mechanism of validation of this classification is carried out, being a subgroup from patterns reinforced as far as probability that the classification made previously has been the correct one, and being ambiguous, or being rejected, another subgroup that hypothetical had been classified erroneously. After this filtering of the patterns classified initially, in the second subphase the new subgroup of patterns classified is used to recover those that are ambiguous. It is important to emphasize the degree of specialization of this classifier: it is based only on a subgroup of patterns of the own form. For that reason, the problem multifont has been reduced to monofont as much to reject incorrect classifications as to recover ambiguous patterns.

In the first subphase, rejecting the patterns with a low evidence of correct classification (ambiguous patterns) will reduce the error rate. In our experiments, 2.05% of the sample patterns were rejected. Nevertheless, this subphase produced 0.3% of misclassifications (patterns not rejected but erroneously classified). Therefore, 97.65% of patterns were classified correctly. Note that, after this subphase, we compensate the slight reduction of the accuracy rate by the fall of the error rate to less than a sixth. In the second subphase an important percentage of most of the rejected patterns are recovered. In our sample, 86% of ambiguous patterns was correctly reclassified, which account for 1.77% of the sample, 0.28% remaining as erroneously reclassified.

The global classification process and the results obtained in our experiments are summarized in Figure 3. When accumulating hit and error rates in both subphases of the specialized classifier, it results in a global accuracy rate of 99,42% with only 0.58% misclassifications. Note that this result fits within our initial goal of 0.7% as the maximum error rate allowed.

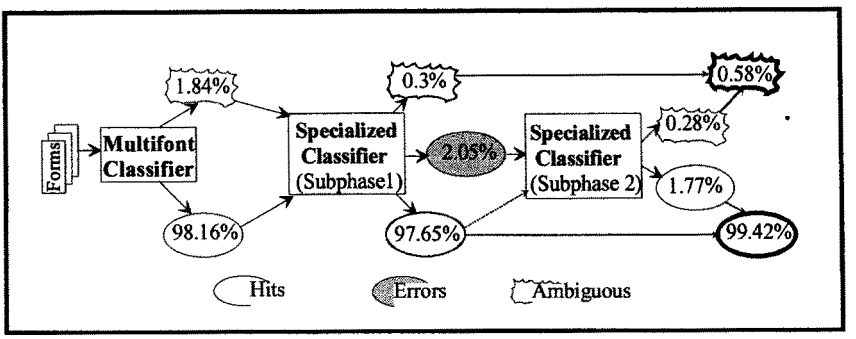

Fig. 3. Classification phases and results.

Another important consequence derived from the introduction of the specialization phase can be seen in the behavior of the distribution of the error count per form. The dispersion of the errors by form after the specialized classification diminishes remarkably. Before this phase, only a 14% of the forms had zero errors, whereas after that, this percentage increases to a 37%. This behavior implies that the manual recognition that there is to make in the end with the digits badly classified is outstandingly going to be simpler (remember that there are near 500 digits per form), increasing of this form the benefits of the system. The low dispersion of errors invites the application of clustering techniques to aid the manual correction of erroneous forms.

We analysed the errors and labelled them according to their nature (broken digits, noise, excessive blurring, etc.). In most cases, the errors were due to deficiencies in previous phases of the recognition process, out of our control, and they can be solved by applying particular techniques, as contour analysis, background analysis, etc.

4 Conclusions

This article has presented a hierarchical classifier to be used in applications of massive form recognition (250,000 forms of about 500 digits each one) characterized to present broken and blurred typewritten digits. We have used a global feature, zoning, mainly by its tolerance to the blurring component.

A classifier in two phases has been designed. First a method 1-NN makes a classification multifont. This method has been selected after the experimentation made with the classifiers presented in Section 3. In addition, the effect of inferior breaks in the digits has been quantified (a 25% of the sample present inferior breaks), introducing a mechanism in the classifier to improve the global accuracy of the system. In the second phase, the high probability of finding just one font in one single form is the foundation for making a monofont classification using a medium distance metric among the classes in the form. In our experiments, the accuracy rate of the system rises after the second phase from 98.16% to 99.42%, reducing the error rate by more than two-thirds.

The performance of the system can be expressed in number of operations that the recognition of a pattern requires in each one of the phases: a) in phase 1, the number of operations by pattern is of the order of 50,000 and b) in phase 2, it is of the order of 20,000. This number of operations assumes in a sequential computer. We are currently making some improvements in two lines: 1) to diminish the number of patterns of reference (at the present time of the order of 1,000) using the technique of diagrams of Voronoi, and 2) to develop a parallel algorithm for doing the classification. Also, our future research is aimed at generalizing the classifier to alphanumeric and handwritten characters.

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 Departament (UPV/EHU), and Pentha Magnetics S.A.; and by Gipuzkoako Foru Aldundia and University of the Basque Country.

6 References

[1] **M.A. Abou-Nasr, M.A. Sid-Ahmed:** Fast Learning and Efficient Memory Utilization with a Prototype Based Neural Classifier. *Pattern Recognition*, Vol. 28, No. 4, pp. 581-593, 1995.

[2] **F.H. Cheng, W.H. Hsu, M.C. Kuo:** Recognition of Handprinted Chinese Characters Via Stroke Relaxation. *Pattern Recognition*, Vol. 26, No. 4, pp. 579-593, 1993.

[3] **Z. Chi, J. Wu, H. Yan:** Handwritten Numeral Recognition Using Self-Organizing Maps and Fuzzy Rules. *Pattern Recognition, Vol. 28, No. 1, pp. 59-66, 1995.*

[4] **B.V. Dasarathy:** *Nearest Neighbor(NN) Norms: NN Pattern Classification Techniques.* Ed: IEEE Computer Society Press, 1991.

[5] **S.A. Dudani, K.J. Breeding, R.B. McGhee:** Aircraft Identification by Moment Invariants. *IEEE Transactions on Computers*, Vol. C-26, No. 1, pp. 39-45, January 1977.

[6] **M.D. Garris, C.L. Wilson, J.L. Blue, G.T. Candela, P. Grother, S. Janet, R.A. Wilkinson:** Massively Parallel Implementation of Character Recognition Systems. *NIST: SPIE's Conference on Character Recognition and Digitizer Technologies*, Vol. 1661, pp. 269-280, February 1992.

[7] **J. Geist, R.A. Wilkinson, S. Janet, P. Grother, B. Hammond, N.W. Larsen, R.M. Klear, M.J. Matsko, C.J.C. Burges, R. Creecy, J.J. Hull, T.P. Vogl, C.L. Wilson:** *The Second Census Optical Character Recognition Systems Conference.* NIST: Technical Report NISTIR 5452, National Institute of Standards Technology, May 1994.

[8] **P. Grother, G.T. Candela:** Comparison of Handprinted Digit Classifiers. *NIST: Technical Report* NISTIR 5209, National Institute of Standards Technology, June 1993.

[9] **S. Impedovo, L. Ottaviano, S. Occhinegro:** Optical Character Recognition - A Survey. Character & Handwriting Recognition, Ed: P.S.P. Wang, pp. 1-24, *World Scientific series in Computer Science,* Vol. 30, 1991.

[10] **S. Knerr, L. Personnaz, G. Dreyfus:** Handwritten Digit Recognition by Neural Networks with Single-Layer Training. *IEEE Transactions on Neural Networks*, Vol. 3, No. 6, pp. 962-968, November 1992.

[11] **R.P. Lippmann:** An Introduction to Computing with Neural Nets. *IEEE ASSP*, pp. 4-22, April 1987.

[12] **O. Matan, H.S. Baird, J. Bromley, C.J.C. Burges, J.S. Denker, L.D. Jackel, Y. Le Cun, E.D.P. Pednault, W.D. Satterfield, C.E. Stenard, T.J. Thompson:** Reading Handwritten Digits: A ZIP Code Recognition System. *IEEE Computer*, pp. 59-63, July 1992.

[13] **D. Michie, D.J. Spiegelhalter, C.C. Taylor:** *Machine Learning, Neural and Statistical Classification.* Ed: Ellis Horwood Series in Artificial Intelligence, 1994.

[14] **S. Mori, C.Y. Suen, K. Yamamoto:** Historical Review of OCR Research and Development. *Proceedings of the IEEE*, Vol. 80, No. 7, pp. 1029-1058, July 1992 *(Special Issue on Optical Character Recognition).*

[15] **J. Muguerza:** *Una Solución al Reconocimiento Automático de Dígitos Imprecisos en Formularios.* Tesis Doctoral, Departamento de Arquitectura y Tecnología de Computadores, Universidad del País Vasco, Enero 1996.

[16] **L. O'Gorman, R. Kasturi:** Document Image Analysis Systems. Introduction. *IEEE Computer*, pp. 5-8, July 1992.

[17] **D.E. Rumelhart, G.E. Hinton, R.J. Williams:** Learning Internal Representations by Error Propagation. D.E. Rumelhart, J. MacClelland (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1: Foundations.* The Massachusetts Institute of Technology, Cambridge, MA, USA, 1986.

[18] **P.K. Simpson:** *Artificial Neural Systems: Foundations, Paradigms, Applications and Implementations.* Ed: Pergamon Press, 1990.

[19] **S.N. Srihari:** High-Performance Reading Machines. *Proceedings of the IEEE*, Vol. 80, No. 7, pp. 1120-1132, July 1992 *(Special Issue on Optical Character Recognition).*

[20] **C.Y. Suen, M. Berthod, S. Mori:** Automatic Recognition of Handprinted Characters– The State of Art. *Proceedings of the IEEE*, Vol. 68, No. 4, pp. 469-487, April 1980.