Hierarchical Classification of Handwritten Characters based on Novel Structural Features

G. Vamvakas , B. Gatos and S.J. Perantonis

Computational Intelligence Laboratory,
Institute of Informatics and Telecommunications,
National Center for Scientific Research “Demokritos”,
GR-153 10 Agia Paraskevi, Athens, Greece
http://www.iit.demokritos.gr/cil,
{gbam,bgat,sper}@iit.demokritos.gr

Abstract

In this paper, we present a methodology for off-line handwritten character/digit recognition. The proposed methodology relies on a new feature extraction technique based on recursive subdivisions of the image as well as on calculation of the centre of masses of each sub-image. Feature extraction is followed by a hierarchical classification scheme based on the level of granularity of the feature extraction method. Pairs of classes with high values in the confusion matrix are merged at a certain level and higher level granularity features are employed for distinguishing them. A handwritten character database as well as a handwritten digit database is used in order to demonstrate the efficiency of the proposed technique.

Keywords: Handwritten character/digit recognition, feature extraction, hierarchical classification.

1. Introduction

A widely used approach in Optical Character Recognition (OCR) systems is to follow a two step schema: a) represent the character as a vector of features and b) classify the feature vector into classes [3]. Selection of a feature extraction method is most important in achieving high recognition performance. A feature extraction algorithm must be robust enough so that for a variety of instances of the same symbol, similar feature sets are generated, thereby making the subsequent classification task less difficult [4].

In the literature, feature extraction methods for handwritten characters and digits have been based on two types of features: a) statistical, derived from statistical distribution of points, b) structural. The most common statistical features used for character representation are: a) zoning, where the character is divided into several zones and features are extracted from the densities in each zone [5] or from measuring the direction of the contour of the character by computing histograms of chain codes in each zone [6], b) projections [7] and c) crossings and distances [8]. Structural features are based on topological and geometrical properties of the character, such as maxima and minima, reference lines, ascenders, descenders, cusps above and below a threshold, cross points, branch points, strokes and their directions, inflection between two points, horizontal curves at top or bottom, etc [9]. A survey on feature extraction methods can be found in [10].

Classification methods on learning from examples have been applied to character recognition mainly since the 1990s. These methods include statistical methods based on Bayes decision rule, Artificial Neural Networks (ANNs), Kernel Methods including Support Vector Machines (SVM) and multiple classifier combination [11], [12].

Most character recognition techniques described in the literature use a “one model fits all” approach, i.e. a set of features and a classification method are developed and every test pattern is subjected to the same process. Some approaches which employ a hierarchical treatment of patterns have also been proposed in the literature. As shown in [13], this approach can have considerable advantages compared to the “one model fits all” approach.

In this paper a novel feature extraction method based on recursive subdivisions of the character image is presented. This feature extraction scheme represents the characters at different levels of granularity. Even though the method is quite efficient when a specific level of granularity is used, we show that more is to be gained in classification accuracy by exploiting the intrinsically recursive nature of the method. This is achieved by appropriately combining the results from different levels using a hierarchical approach. Lower levels are used to perform a preliminary discrimination, whereas higher levels help in distinguishing between characters of similar shapes that are confused when using only lower levels. The remaining of this paper is organized as follows. In Section 2 the proposed OCR methodology is presented while experimental results are discussed in Section 3. Finally, conclusions are drawn in Section 4.

2. OCR Methodology

The proposed OCR methodology follows a two step schema: First a feature extraction method is applied to obtain the feature vectors and then a hierarchical classification scheme is performed.

2.1. Feature Extraction

In this section a new feature extraction method for handwritten character recognition is presented. This method is based on structural features, extracted directly from the character image, that provide a good representation of the character at different levels of granularity.

Let $im(x,y)$ be the character image array having 1s for foreground and 0s for background pixels and x_{max} and y_{max} be the width and the height of the character image. Our feature extraction method is based on the centre of mass of the character image. First, the co-ordinates (x_o, y_o) of the centre of mass of the initial character image are calculated. In order to avoid quantizing errors, we use the following equations:

$$x_o = \arg\min_{x_t} \left\{ \sum_{\substack{x=1 \\ y=1}}^{\substack{x=x_t \\ y=y_{\max}}} im(x,y) - \sum_{\substack{x=x_t+1 \\ y=1}}^{\substack{x=x_{\max} \\ y=y_{\max}}} im(x,y) \right\} \qquad (1)$$

$$y_o = \arg\min_{y_t} \left\{ \sum_{\substack{x=1 \\ y=1}}^{\substack{x=x_{\max} \\ y=y_t}} im(x,y) - \sum_{\substack{x=1 \\ y=y_t+1}}^{\substack{x=x_{\max} \\ y=y_{\max}}} im(x,y) \right\} \qquad (2)$$

These co-ordinates divide the image into four rectangular sub-images with the following upper left and lower right vertex coordinates:

$$\{(1,1),(x_o,y_o)\} \ , \ \{(x_o,1),(x_{\max},y_o)\}$$
$$\{(1,y_o),(x_o,y_{\max})\} \ , \ \{(x_o,y_o),(x_{\max},y_{\max})\}$$

with each one consisting of almost the same amount of foreground pixels. This procedure is applied recursively for every sub-image (see Fig.1)

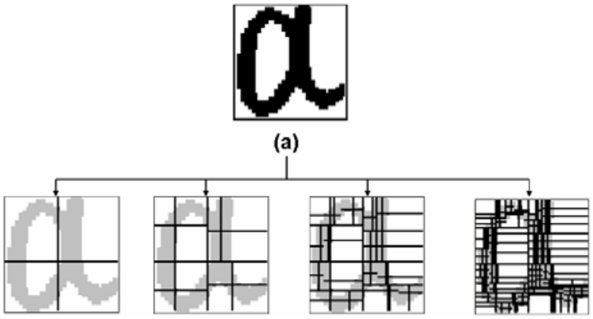

Figure 1. Character image and sub-images based on centre of mass: (a) original image, (b), (c), (d), (e) subdivisions at levels 0, 1, 2 and 3 respectively.

Let L be the current level of the granularity. At this level the number of the sub-images is $4^{(L+1)}$. For example, when L=0 (Fig.1b) the number of sub-images is 4 and when L=1 it is 16 (Fig.1c). The number of the center of masses at level L equals to 4^L (see Fig.2). At level L, the co-ordinates (x,y) of all the centre of masses are stored as features. So, for every L a $2*4^L$ - dimensional feature vector is extracted. As Fig.2 shows, the larger the L the better representation of the character is obtained.

Finally, after all feature vectors are extracted each feature is normalized to [0, 1]. Let m_i be the mean value of the i_{th} feature for all training vectors and σ_i the standard deviation respectively. Then the value f_i of the i_{th} feature of every feature vector is normalized according to Eq.3.

$$f_i^{'} = \frac{\dfrac{f_i - m_i}{3\sigma_t} + 1}{2} \qquad (3)$$

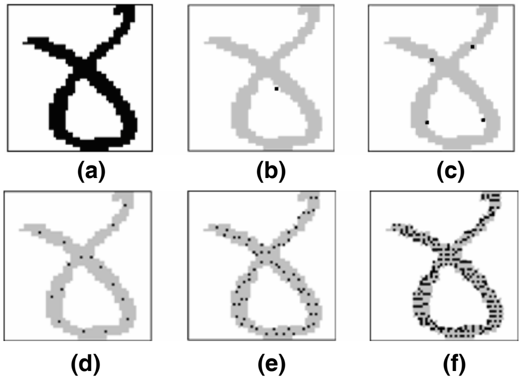

Figure 2. Features based on centre of mass: (a) original image, (b), (c), (d), (e), (f) features at levels 0, 1, 2, 3 and 4 respectively.

2.2. Hierarchical Classification

For the recognition stage a hierarchical classification scheme is employed. Since characters with similar shapes i.e. 'ξ' and 'ζ' from the Greek alphabet, are often mutually confused when using a low granularity feature representation, we propose to merge the corresponding classes to the certain level of classification. At a next step, we distinguish those character classes by employing a higher granularity feature extraction vector at a hierarchical classification scheme. The hierarchical classification scheme has five distinct steps and the flowchart of the whole procedure is shown in Fig.3.

Step 1: Extract features at levels 1, 2, 3 and 4. These levels are considered to be the most appropriate for feature extraction since level 0 is just one point (the centre of mass of the initial image) that has no meaningful information and levels above 4 tend to extract features that depend on

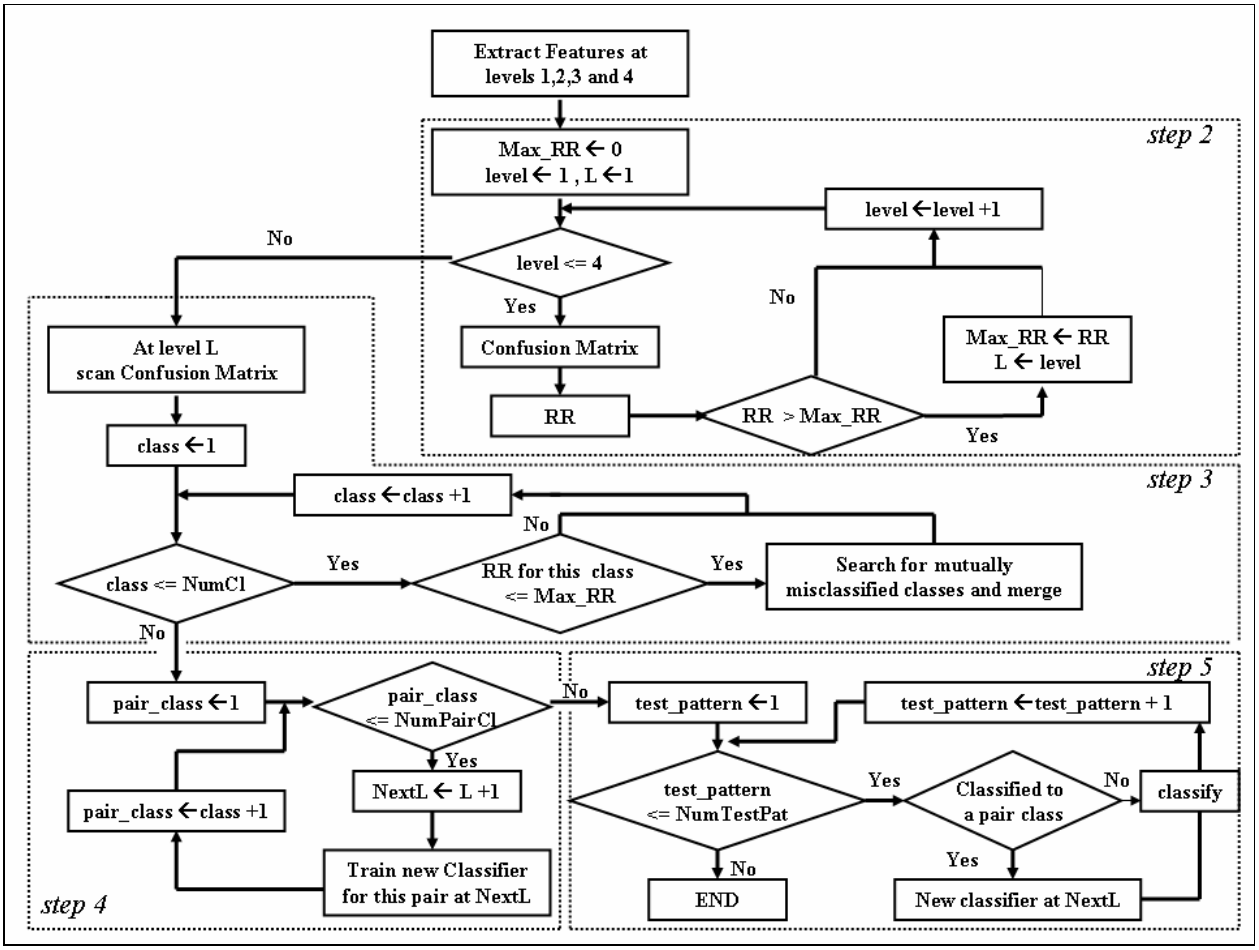

Figure 3. Flowchart of the recognition procedure: Max_RR = Maximum Recognition Rate, RR= Recognition Rate, L = level at which we have the maximum recognition rate, NumCl = Total Number of Classes, NumPairCl= Total Number of Pair Classes, NumTestPat = Total Number of Test Patterns.

the specific character each time thus the dimensionality of the feature vectors is too large.

Step 2: For features extracted at step 1 construct the confusion matrices from the training set using a *K*-fold cross-validation process. The training set is divided into *K* subsets such that the analysis is initially performed on a single subset, while the other subsets are retained for subsequent use in confirming and validating the initial analysis. In our case *K* is set to 10. From these confusion matrices calculate the overall recognition rates at each level. Features from level *L* with the highest recognition rate are considered to be the initial features used for the classification procedure.

Step 3: Let the overall recognition rate among all categories for the best performing level *L* of granularity be a threshold. At this level *L*, the confusion matrix is scanned and classes whose recognition rate is below the threshold are detected. For each one of these classes find the class with which they are mutually misclassified the most and consider them to be one pair.

Step 4: For each one of the pair classes found in Step 3 another classifier is trained with features extracted at level *L* + 1 of the granularity procedure in order to distinguish them at a later stage of the classification.

Step 5: Each pattern of the test set is then fed to the initial classifier with features extracted at level *L*. If the classifier decides that this pattern belongs to one of the non-pair classes then its decision is taken into account and the unknown pattern is assumed to be classified. Else, if it is classified to one of the pair classes then it is given to the pair's corresponding classifier and this new classifier decides the recognition result.

3. Experimental Results

For our experiments the CIL Database [1] (Fig.4) and the MNIST Database [2] (Fig.5) were used. In the particular recognition problem, the classification step was performed using SVM [14] with Radial Basis Function (RBF) kernel.

The CIL database comprises samples of 46 Greek written by 125 Greek writers. Every writer contributed 5 samples of each letter, thus resulting to a database of 625 variations per letter and an overall of 28,750 isolated labeled characters. Each character is normalized to a *N*x*N* matrix. For our experiments *N*=60. Moreover, 1/5 of each class was used for testing and the remaining 4/5 for training.

The MNIST Database consists of 70,000 isolated and labeled handwritten digits. It is divided into a training set of 60,000 and a test set of 10,000 digits.

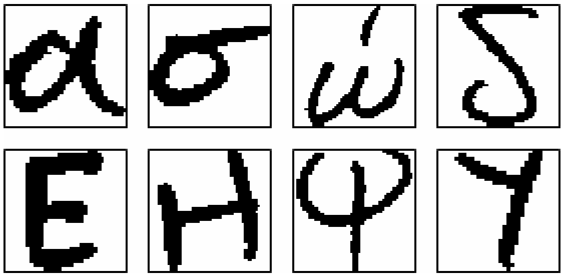

Figure 4. Samples from the CIL Database.

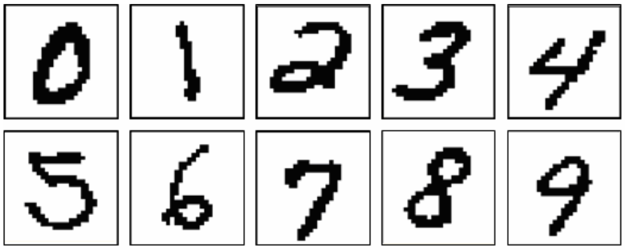

Figure 5. Samples from the MNIST Database.

As described in Section 2, first features at levels 1, 2, 3 and 4 are extracted for all patterns in training set and the confusion matrices at each level are constructed. As shown in Table 1, for the CIL database the highest overall recognition rate (92.10%) is achieved when features from level 3 are used. In Table 2, we present a comparison of this result with other state-of-the-art feature extraction methods for handwritten character recognition. These methods are the following:

(a) A hybrid feature extraction scheme based on zones and projections (HYB) [15]
(b) A scheme based on structural features (STR) [15]
(c) Features based on both statistical and structural methods with a dimensionality reduction scheme (DIM) [1]

From Table 2, it is evident that using the proposed features we obtain a recognition rate which is at par with the best state-of-the-art techniques.

Next, the results of the hierarchical approach are presented. Since the best recognition is achieved in level 3, features obtained at this level are used to train the initial SVM. Then, the confusion matrix at level 3 is scanned and for every class whose recognition rate is below 92.10% the class with which is mutually misclassified the most is detected. Table 3 shows the most confused pairs of classes. Each pair is merged into one class and for every pair a new SVM is trained with features from level 4 in order to distinguish them at a next stage. As shown in the last row of Table 2, when the hierarchical classification scheme is used the overall recognition rate is improved (93.21%).

Table 1. Recognition rates using CIL Database.

CIL Database	
Level 1	65.91%
Level 2	91.01%
Level 3	92.10%
Level 4	90.90%

Table 2. Comparison of the proposed OCR methodology.

CIL Database	
HYB[15]	91.61%
STR [15]	88.62%
DIM [1]	92.05%
Proposed methodology (Level 3)	**92.10%**
Proposed methodology (Hierarchical Classification)	**93.21 %**

Table 3. Mutually misclassified classes for features at level 3 for CIL Database.

Class 1	Class 2
β	Β
ζ	Ξ
ι	Ι
ν	Υ
τ	Ζ
φ	Ψ
έ	Ί
Α	Λ

Regarding the MNIST database, features were again extracted at levels 1, 2, 3 and 4 and four confusion matrices were calculated respectively. Again, recognition rate was higher when using features from level 3 (see Table 4). Then, mutually misclassified classes were detected (see Table 5) and for each pair a new SVM was trained with features from level 4. As shown in Table 4, when the hierarchical classification scheme is used the overall recognition rate is improved (98.66%).

Table 4. Recognition rates using MNIST Database.

MNIST Database	
Level 1	80.87%
Level 2	96.41%
Level 3	97.78%
Level 4	97.15%
Hierarchical Classification	**98.66%**

Table 5. Mutually misclassified classes for features at level 3 for MNIST Database.

Class 1	Class 2
3	8
4	9
5	6

4. Conclusions

In this paper a novel feature extraction method for handwritten characters and digits was presented based on recursive subdivisions of the character image. This feature extraction scheme represents the characters at different levels of granularity. Even though the method is quite efficient when a specific level of granularity is used, we show that more is to be gained in classification accuracy by exploiting the intrinsically recursive nature of the method. This is achieved by appropriately combining the results from different levels using a hierarchical approach. Lower levels are used to perform a preliminary discrimination, whereas higher levels help in distinguishing between characters of similar shapes that are confused when using only lower levels. As shown at the experimental results, the proposed hierarchical classification scheme outperforms other state-of-the-art feature extraction techniques.

References

[1] G. Vamvakas, B. Gatos, S. Petridis and N. Stamatopoulos, "An Efficient Feature Extraction and Dimensionality Reduction Scheme for Isolated Greek Handwritten Character Recognition", *Proceedings of the 9th International Conference on Document Analysis and Recognition*, Curitiba, Brazil, 2007, pp. 1073-1077.

[2] http://yann.lecun.com/exdb/mnist/

[3] A. S. Brito, R. Sabourin, F. Bortolozzi, "Foreground and Background Inforamtion in a HMM-Based Method for Recognition of Isolated Characters and Numeral Strings", *Proceedings of the 9th International Workshop on Frontiers in Handwritten Recognition*, 2004, pp. 371-376.

[4] J. A. Fitzgerald, F. Geiselbrechtinger, and T. Kechadi, "Application of Fuzzy Logic to Online Recognition of Handwritten Symbols", *Proceedings of the 9th International Workshop on Frontiers in Handwritten Recognition*, 2004, pp. 395-400.

[5] Luiz S. Oliveira, F. Bortolozzi, C.Y.Suen, " Automatic Recognition of Handwritten Numerical Strings: A Recognition and Verification Strategy", *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 2001, Vol. 24, No. 11, pp. 1448-1456.

[6] K. M. Mohiuddin and J. Mao, "A Comprehensive Study of Different Classifiers for Handprinted Character Recognition", *Pattern Recognition*, Practice IV, 1994, pp. 437- 448.

[7] A. L. Koerich, "Unconstrained Handwritten Character Recognition Using Different Classification Strategies", *International Workshop on Artificial Neural Networks in Pattern Recognition* (ANNPR), 2003.

[8] J. H. Kim, K. K. Kim, C. Y. Suen, " Hybrid Schemes Of Homogeneous and Heterogeneous Classifiers for Cursive Word Recognition", *Proceedings of the 7th International Workshop on Frontiers in Handwritten Recognition,* Amsterdam, 2000, pp 433 - 442.

[9] N. Arica and F. Yarman-Vural, " An Overview of Character Recognition Focused on Off-line Handwriting ", *IEEE Transactions on Systems, Man, and Cybernetics*, Part C: Applications and Reviews, 2001, 31(2), pp. 216 - 233.

[10] O. D. Trier, A. K. Jain, T.Taxt, "Features Extraction Methods for Character Recognition – A Survey ", *Pattern Recognition*, 1996, Vol.29, No.4, pp. 641-662.

[11] C. L. Liu, H. Fujisawa, "Classification and Learning for Character Recognition: Comparison of Methods and Remaining Problems", *Int. Workshop on Neural Networks and Learning in Document Analysis and Recognition,* Seoul, 2005.

[12] F. Bortolozzi, A. S. Brito, Luiz S. Oliveira and M. Morita, "Recent Advances in Handwritten Recognition", *Document Analysis,* Umapada Pal, Swapan K. Parui, Bidyut B. Chaudhuri, pp 1-30.

[13] J. Park, V. Govindaraju, S. N. Shrihari, "OCR in Hierarchical Feature Space", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, Vol. 22, No. 24, pp. 400-408.

[14] Cortes C., and Vapnik, V, " Support-vector network ", *Machine Learning*, vol. 20, pp. 273-297, 1997.

[15] G. Vamvakas, N. Stamatopoulos, B. Gatos, I. Pratikakis and S. J. Perantonis, "Greek Handwritten Character Recognition", *Proceedings of the 11th Panhellenic Conference in Informatics*, 18-20 May 2007, Patras, Greece. Vol.B, pp 343- 352.