A Novel Feature Extraction and Classification Methodology for the Recognition of Historical Documents

G. Vamvakas, B. Gatos and S. J. Perantonis
Computational Intelligence Laboratory , Institute of Informatics and Telecommunications,
National Centre for Scientific Research "Demokritos",
GR – 153 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 character recognition that mainly focuses on handling the difficult cases of historical fonts and styles. 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 with sub-pixel accuracy. 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. Several historical documents were used in order to demonstrate the efficiency of the proposed technique.

1. Introduction

Nowadays, the recognition of both contemporary machine-printed and isolated handwritten characters is performed with high accuracy. However, the recognition of historical documents still remains an open problem in the research arena due to the difficult cases of historical fonts and styles. 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of a feature extraction method is most important in achieving high recognition performance especially in the case of historical documents where we have a large number of different symbols and styles. 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 work, a dynamic character recognizer is presented. The recognizer begins with features extracted in a coarse resolution and focuses on smaller sub-images of the

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.223

character on each recursive pass, thus working with a finer resolution of a sub-image each time, till classification meets acceptance criteria.

In this paper we present a novel feature extraction method based on recursive subdivisions of the character image as well as on calculation of the centre of masses of each sub-image with sub-pixel accuracy. 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 session a new feature extraction method, for the recognition of machine printed and handwritten historical documents, 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 and permit handling the difficult cases of historical fonts and styles.

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 relies on recursive sub-divisions of the character image based on the centre of mass of each sub-image. In order to avoid quantizing errors and improve the precision, the centres of masses are calculated with sub-pixel accuracy. First, the co-ordinates (x_o, y_o) of the centre of mass of the initial character image are calculated. The vertical co-ordinate x_0 is found according to the following procedure:

Step 1: Let V_o be the vertical projection array of the initial character image.

Step 2: Create V_1 array from V_0 as follows:

```
for x = 1 to 2 * $x_{max}$
    if x mod 2 = 1 then
        $V_1[x] = 0$
    else
        $V_1[x] = V_0[x$ div $2]$
    end if
end for
```

Step 3: Find x_q from V_1 using the following equation:

$$x_q = \arg\min_{x_t} \left\{ \sum_{x=1}^{x=x_t-1} V_1(x) - \sum_{x=x_t+1}^{x=2*x_{\max}} V_1(x) \right\} \tag{1}$$

Step 4: The vertical co-ordinate x_0 is then estimated as:

$$x_o = x_q div 2 \tag{2}$$

As already mentioned, in order to improve the precision, the centre of mass for each of the following sub-images is calculated with sub-pixel accuracy. That is, the initial image is divided vertically into two rectangular sub-images depending on the value of x_q (Eq 1). If $x_q \ mod \ 2 = 0$ then the vertex co-ordinates of these two sub-images are: {(1, 1), (x_0, y_{max})} and {(x_0, 1), (x_{max}, y_{max})}. Otherwise, if $x_q \ mod \ 2 = 1,$ then the vertex co-ordinates are: {(1, 1), (x_0, y_{max})} and {(x_0 +1, 1), (x_{max}, y_{max})}.

Likewise, the horizontal co-ordinate y_0 is calculated thus resulting to the division of the initial character images into four rectangular sub-images. The whole procedure is applied recursively for every sub-image (Fig.1).

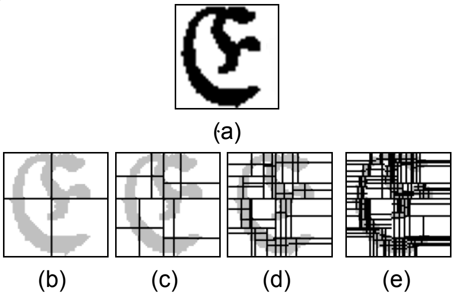

Figure 1. Character image and sub-images based on centre of mass: (a) original image, (b), (c), (d), (e) subdivisions at levels 0, 1, 2 and 3 respectively.

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 (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. Up to here two questions rise as one can easily realize. First, at which level L of granularity the best recognition result is achieved and second, how deep the penetration will be. Both questions are answered at the following section of the paper.

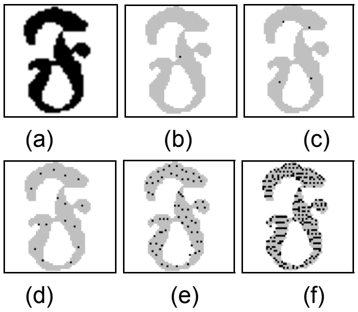

Figure 2. Features based on centre of mass: (a) original image, (b), (c), (d), (e), (f) features at levels 0, 1, 2, 3 and 4 respectively.

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.

$$f_i' = \frac{(f_i - m_i)/_{3\sigma_i} + 1}{2} \tag{3}$$

2.2. Hierarchical Classification

For the recognition stage a hierarchical classification scheme is employed. Since characters with similar structure i.e. 'ζ' and 'ξ' 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 four distinct steps; three for training phase and one for testing.

Step 1: Starting from level 1 and gradually proceeding to higher levels of granularity features are extracted, the confusion matrix is created and the overall recognition rate is calculated, until the recognition rate stops increasing (Fig.3). Features from level L with the highest recognition rate (*Max_RR*) are considered to be the initial features used for the classification procedure. Confusion matrices are created at each level from the training set using a *K*-fold cross-validation process. In *K*-fold cross-validation, the original training set is partitioned into *K* subsets. Of the *K* subsets, a single subset is retained as the validation data for testing the model, and the remaining *K*−1 subsets are used as training data. The cross-validation process is then repeated *K* times (the *folds*), with each of the *K* subsets used exactly once as the validation data. The *K* results from the folds then can be averaged (or otherwise combined) to produce a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. In our case *K* is set to 10.

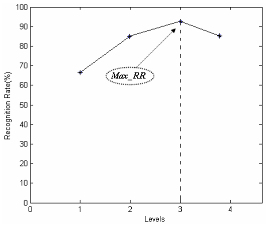

Figure 3. Example of finding the Level *L* of granularity with the highest recognition rate.

Step 2: Let the overall recognition rate among all categories for the best performing level L of granularity, that is *Max_RR*, be a threshold. At this level L, the corresponding confusion matrix is scanned and classes whose recognition rate is below this 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 3: For each one of the pair classes found in Step 2 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 4: 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 two databases comprising samples of characters from old Greek Christian documents of the 17th century and the CEDAR CD-

ROM1 [1] database were used. In the particular classification problem, classification step was performed using SVM [14] with Radial Basis Function (RBF).

In [15] two databases using a clustering scheme, are created from old Greek Christian documents. A typewritten (*TW*) consisting of 13,966 characters from 67 classes and a handwritten (*HW*) one of 6,758 characters from 51 classes. Moreover for both databases, the 80% of each class is used for training while the rest 20% is used for testing (Table 1).

Table 1. The historical typewritten (*TW*) and handwritten (*HW*) databases.

Database	Data Set	Number of Classes	Train Set	Test Set
TW	13,966	67	11,173	2,793
HW	6,758	51	5,407	1,351

The CEDAR database consists of samples of 52 English handwritten characters: 19,145 characters were used for training and 2,183 characters for testing.

For all three databases each character is normalized to an *N*x*N* matrix. In our case, *N* = 60.

According to Section 2, the best performing level is first found. As shown in Table 2, for the *TW-Database* the best performing level (97.59%) is 2. Then, the confusion matrix at level 2 is scanned and for every class whose recognition rate is below 97.59% the class with which it 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 3 in order to distinguish them at a next stage. Table 2 depicts the recognition rates achieved at each level and the recognition rate using the hierarchical classification procedure.

Table 2. Recognition rates using the TW-Database.

TW - Database	
Level 1	91.46%
Level 2	97.59%
Level 3	95.55%
Hierarchical Classification	**97.71 %**

Table 3. Mutually misclassified classes for features at level 2 for the TW - Database.

Class 1	Class 2
α	ο
ς	ϛ (στ)
ν	υ
τ	Τ
Α	Λ

For the *HW-Database*, the highest recognition rate (93.14%) is achieved when features from level 3 are used (Table 4), mutually misclassified classes at this level are found (Table 5) and again the overall recognition rate is improved when the hierarchical classification scheme is applied.

Table 4. Recognition rates using the HW-Database.

HW - Database	
Level 1	81.50%
Level 2	91.96%
Level 3	93.14%
Level 4	89.31%
Hierarchical Classification	**94.51 %**

Table 5. Mutually misclassified classes for features at level 3 for the HW - Database.

Class 1	Class 2
κ	ϗ (και)
α	ο
ρ	φ
ε	Σ

In [15], an evaluation of these two databases is also presented. In order to do so, a hybrid feature extraction scheme is employed based on zones and projections combined in a fusion way. The comparison of this methodology with the one proposed in this paper is shown in Table 6. From this table, it is evident that although the recognition rate is almost the same as far as the handwritten database is concerned, when it comes to typewritten characters the improvement is considerably noticeable.

Table 6. Comparison of the proposed OCR methodology for historical characters.

	TW-Database	HW-Database
Hybrid [15]	95.44%	94.62%
Proposed Methodology	97.71%	94.51%

Finally, Tables 7 depicts the recognition results of the proposed methodology when applied to modern characters as well. As mentioned before, for this experiment the CEDAR database was used, trying to distinguish all 52 characters (classes). Table 8 shows the comparison with other state-of-the-art techniques that deal with 52 classes.

4. Conclusions

In this paper a new feature extraction method, for historical machine printed and handwritten characters, was presented based on recursive subdivisions of the character image. As shown at the experimental results, the proposed hierarchical classification scheme outperforms other state-of-the-art feature extraction techniques as far as recognition of historical fonts and styles is concerned while works efficiently enough for contemporary handwritten documents.

Table 7. Recognition rates using the CEDAR Database.

CEDAR Database	
Level 1	56.57%
Level 2	77.28%
Level 3	77.51%
Level 4	75.36%
Hierarchical Classification	**80.19%**

Table 8. Comparison of the proposed OCR methodology for modern characters.

	CEDAR Database
Yamada [16]	75.70%
Kimura [17]	73.25%
Gader [18]	74.77%
Proposed Methodology	**80.19%**

5. Acknowledgement

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement n° 215064 (project IMPACT).

6. References

[1] J.J. Hull, “A database for handwritten text recognition research”, *IEEE Trans. Pattern Anal. Mach. Intell.* 16 (5) (1994) 550–554.

[2] The MNIST Database, http://yann.lecun.com/exdb/mnist/

[3] A. S. Brito, R. Sabourin, F. Bortolozzi, “Foreground and Background Information 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, “Feature 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, B. Gatos, N. Stamatopoulos and S.J.Pernantonis, “A Complete Optical Character Recognition Methodology for Historical Documents”, *Document Abalysis Systems (DAS’08)*, Nara, Japan, 2008, pp.525-532.

[16] H. Yamada and Y. Nakano, “Cursive Handwritten Word Recognition Using Multiple Segmentation Determined by Contour Analysis”, *IECIE Transactions on Information and System*, Vol. E79-D. pp. 464-470, 1996.

[17] F. Kimura. N. Kayahara. Y. Miyake and M. Shridhar, "Machine and Human Recognition of Segmented Characters from Handwritten Words", *International Conference on Document Analysis and Recognition (ICDAR '97),* Ulm, Germany, 1997, pp. 866-869.

[18]. P. D. Gader, M. Mohamed and J-H. Chiang. "Handwritten Word Recognition with Character and Inter-Character Neural Networks*", IEEE Transactions on System, Man. and Cybernetics-Part B: Cybernetics*, Vol. 27, 1997, pp. 158-164.