A-Survey of Feature Extraction and Classification Techniques in OCR Systems

Rohit Verma
School of Information Technology, Apeejay Institute of Management Technical Campus, Jalandhar, India

Dr. Jahid Ali
Sri Sai Iqbal College of Management & Information Technology , Badhani, Pathankot, India

Abstract: - This paper describes a set of feature extraction and classification techniques, which play very important role in the recognition of characters. Feature extraction provides us methods with the help of which we can identify characters uniquely and with high degree of accuracy. Feature extraction helps us to find the shape contained in the pattern. Although a number of techniques are available for feature extraction and classification, but the choice of an excellent technique decides the degree of accuracy of recognition. A lot of research has been done in this field and new techniques of extraction and classification has been developed. The objective of this paper is to review these techniques, so that the set of these techniques can be appreciated.

Keywords: OCR, Features Extraction.

1. Introduction

Optical Character Recognition is a most important gift given by computer science to the mankind. It has made a lot of tedious work easy and speedy. OCR means a technique of recognition of machine printed or hand written text by computer and then its conversion to an editable form as per the requirement. The various phases of OCR technique are:

a) Digitization
b) Pre-processing
c) Segmentation
d) Feature Extraction & Classification
e) Post processing

In the Digitization phase the printed or hand written data is converted into the digital form either by scanning the given document or by writing with a digitizer connected with an LCD. The second phase is the Pre-processing phase. In this phase the basic operations performed are: Binarization, Contour Smoothing, Noise reduction, Skew Detection and finally Skeletonization of the digital image. In the segmentation phase the image of sequence of characters is decomposed into lines, words and then characters. The feature extraction phase analyses the given text segment and selects a set of features that can be used to uniquely identify the given text segment. In the classification section of this phase, features extracted in the feature extraction section are used to identify the text on the basis of pre-decided rules. The output of the OCR system generally contains errors; the debugging responsibility is performed by the last phase that is Post processing phase. Thus all phases of the OCR system are very important, but the most important phase which is responsible for character recognition is **Feature Extraction and Classification**.

2. Feature Extraction

Different characters have different features; on the basis of these features the characters are recognized. Thus Feature extraction can be defined as the process of extracting differentiating features from the matrices of digitized characters. A number of features have been found in literature on the basis of which the OCR system works to recognize the characters. According to C. Y. Suen (1986), Features of a character can be classified into two classes: Global or statistical features and Structural or topological features.

Global or Statistical Features

Global features are obtained from the arrangement of points constituting the character matrix. These features can be easily detected as compared to topological features. Global features are not affected too much by noise or distortions as compared to topological features. A number of techniques are used for feature extraction; some of these are: moments, zoning, projection histograms, n-tuples, crossings and distances.

i) Moments

In this case the moments of different points present in a character are utilized as a feature. Heutte et. al (1998) said that these are most commonly used methods in character recognition . M K Hu (1962) brought the concept of classical moment invariants. Hu's other moments are statistical measure of the pixel distribution about the centre of gravity of the character. S. S. Reddi (1981) proposed Radial and angular moments where as Zernike moments were proposed by Teh and Chin (1988). Similarly Suen et. al. (1980) suggested that characters can also be identified by considering the distance of various pixels present in the character matrix, from the centroid of the character. These are called central moments.

ii) Zoning According to this technique the character matrix is divided into small portions or zones. The densities of pixels in each zone are calculated and used as features. This concept was suggested by Hussain et al (1972).

iii) Projection histograms Projection histograms give us the number of black pixels in the vertical and horizontal directions of the specified area of the character. This concept was introduced by M. H. Glauberman (1956) in a hardware OCR system. Projection histograms may be vertical, horizontal, left diagonal or right diagonal.

iv) n-tuples: According to this method the position of black or white pixels in a character image is considered as a feature. The n-tuple method has been developed by Tarling and Rohwer (1993). This method provides a number of important properties of pixels.

v) Crossings and distances: Some researchers have obtained features by analyzing the counts the character image is crossed by vectors in certain directions or at certain angles.

2.2 Structural or Topological features

These features are related to the geometry of the character set to be considered. Some of these features are concavities and convexities in the characters, number of end points, no of holes in the characters etc. A lot of research has been done by different researchers to find different structural features.

A method for the recognition of multi font printed characters was proposed by Rocha and Pavlidis (1994). For this method structural features like singular points, convex arcs and strokes, have been used. The singular point may be branch point, an ending point, or convex vertex and a sharp corner.

Similarly another method was purposed by A Amin (2000). Here a number of structural features like number of sub words, number of peaks of each sub word, number of loops of each peak, number and position of complimentary characters, the height and width of each peak for recognition of printed Arabic text were used.

Kahan et al. (1987) have developed a structural feature set for recognition of printed text of different font and size. This feature set includes the following information for a character: location and number of holes in the characters, concavities in the skeletal structure, crossings of strokes, vertical end points of the character and bounding box of the character.

3. Classification

Classification basically decides the feature space to which the unknown pattern belongs. It is another most important component of OCR system. Classification is usually done by comparing the feature vectors corresponding to the input character with the representative of each character class. But before doing this the classifier should posses a number of training patterns. A number of classification methods were purposed by different researchers some of these are statistical methods, syntactic methods, template matching, artificial neural networks, kernel methods.

3.1 Statistical methods

The purpose of the statistical methods is to determine to which category the given pattern belongs. By making observations and measurement processes, a set of numbers is prepared, which is used to prepare a measurement vector. Statistical classifiers are automatically trainable. The *k*-NN rule is a non parametric recognition method. This method compares an unknown pattern to a set of patterns that have been already labeled with class identities in the training stage. A pattern is identified to be of the class of pattern, to which it has the closest distance. Another common statistical method is to use Bayesian classification. A Bayesian classifier assigns a pattern to a class with the maximum a posteriori probability. Besides these methods other statistical methods are:

Quadratic Discriminant Function (QDF), Linear Discriminant Function (LDF), Euclidean distance, cross correlation, Mahanalobis distance, Regularized Discriminant Analysis (RDA)

3.2 Syntactic or structural methods

Syntactic methods are good for classifying hand written texts. This type of classifier, classifies the input patterns on the basis of components of the characters and the relationship among these components. Firstly the primitives of the character are identified and then strings of the primitives are checked on the basis of pre-decided rules.

Generally a character is represented as a production rules structure, whose left-hand side represents character labels and whose right-hand side represents string of primitives. The right-hand side of rules is compared to the string of primitives extracted from a word. So classifying a character means finding a path to a leaf.

3.3 Template matching

This is one of the simplest approaches to patter recognition. In this approach a prototype of the pattern that is to be recognized is available. Now the given pattern that is to be recognized is compared with the stored patterns. The size and style of the patterns is ignored while matching.

According to Bortolozzi et al. (2005) the matching techniques can be classified as: Deformable templates & elastic matching, relaxation matching, direct matching.

3.4 Artificial neural networks

A neural networks composed of inter connected elements called neurons. A neural network can trained itself automatically on the basis of examples and efficient tools for learning large databases. This approach is non-algorithmic and is trainable. The most commonly used

family of neural networks for pattern classification task is the feed-forward network, which includes multilayer perception and Radial-Basis Function (RBF) networks. The other neural networks used for classification purpose are Convolutional Neural Network, Vector Quantization (VQ) networks, auto-association networks, Learning Vector Quantization (LVQ). But the limitation of the systems based on neural networks is their poor capability for generality.

3.5 Kernel methods

Some of the most important Kernel methods are Support Vector Machines , Kernel Principal Component Analysis (KPCA), Kernel Fisher Discriminant Analysis (KFDA) etc. Support vector machines (SVM) are a group of supervised learning methods that can be applied to classification. In a classification task usually data is divided into training and testing sets. The aim of SVM is to produce a model, which predicts the target values of the test data. Different types of kernel functions of SVM are: Linear kernel, Polynomial kernel, Gaussian Radial Basis Function (RBF) and Sigmoid

4 Conclusions

In this paper, we have presented a survey of feature extraction and classification techniques for optical character recognition of general scripts. A lot of research has been done in this field. Still the work is going on to improve the accuracy of feature extraction and classification techniques. However the different methods of feature extraction and classification discussed here are very effective and useful for new researchers.

5. References:

1. A. Amin, "Recognition of printed Arabic text based on global features and decision tree learning techniques", *Pattern Recognition*, Vol. 33, pp. 1309-1323, 2000.
2. A. B. S. Hussain, G. T. Toussaint and R. W. Donaldson, "Results obtained using a simple character recognition procedure on Munson's handprinted data", IEEE Transactions on Computers, pp. 201-205, 1972.
3. C. H. Teh and R. T. Chin, "On image analysis by the method of moments", IEEE Transactions on PAMI, Vol. 10(4), pp. 496-513, 1988.
4. C. Y. Suen, "Character recognition by computer and applications", in Handbook of Pattern Recognition and Image Processing, New York: Academic, pp. 569-586, 1986.
5. C. Y. Suen, M. Berthod and S. Mori, "Automatic recognition of hand printed characters- the state of the art", Proceedings of the IEEE, Vol. 68(4), pp. 469-487,
1980.
6. F. Bortolozzi, A. Britto Jr., L. S.Oliveria and M. Morita, "Recent advances in
handwriting recognition", in the Proceedings of International Workshop on Document Analysis (IWDA), India, pp. 1-30, 2005.
7. J. Rocha and T. Pavlidis, "A shape analysis model with applications to a character recognition system", IEEE Transactions on PAMI, Vol. 16(4), pp. 393-404, 1994.
8. L. Heutte, T. Paquet, J. V. Moreau, Y. Lecourtier and C. Olivier, structural/statistical feature based vector for handwritten character recognition", Pattern Recognition Letters, Vol. 19(7), pp. 629-641, 1998.
9. M. H. Glauberman, "Character recognition for business machines", Electronics, Vol.29, pp. 132-136, 1956.
10. M. K. Hu, "Visual pattern recognition by moment invariants", IRE Transactions on Information Theory, Vol. 8(2), pp. 179-187, 1962.
11. R. Tarling and R. Rohwer, "Efficient use of training data in the n-tuple recognition method", Electronics Letters, Vol. 29(24), pp. 2093-2094, 1993.
12. S. Kahan, T. Pavlidis and H. S. Baird, "On the recognition of printed characters of any font and size", IEEE Transactions on PAMI, Vol. 9(2), pp. 274-288, 1987.
13. S. S. Reddi, "Radial and angular moment invariants for image identification", IEEE Transactions on PAMI, Vol. 3(2), pp. 240-242, 1981.