ISSN:0975-9646

Optical Character Recognition for English and Tamil Script

Kanimozhi.V.M*[1], Muthumani.I*[2]

*[1]PG student *[2]Professor
Department of Electronics and Communication Engineering
Alagappa Chettiar College of Engineering & Technology, Karaikudi, India.

Abstract— optical character recognition is an evergreen area of research and is verily used in various real time applications. This paper proposes a new technique of optical character recognition using Hough transform and statistical method. This method proves to be very effective with the use of randomized Hough transform for feature extraction and Statistical method for developing the model. The model proposed is trained and validated for two languages- cursive English and Tamil script and the results are found to be very much encouraging. The model developed works for the entire character set in both the languages.

Keywords— Hough transform, optical character recognition, statistical method

I. INTRODUCTION

Character recognition has a great potential in data and word processing for instance, automated postal address and ZIP code reading, data acquisition in bank checks, processing of archieved institutional records, etc. In the Recent years, optical character recognition(OCR) has gained a momentum since the need for converting the scanned images into computer recognizable formats such as text documents has increased applications. OCR is one of the most fascinating and challenging areas of pattern recognition with various practical applications.

The process of character recognition involves extraction of defined characteristics called features to classify an unknown character into one of the known classes. Therefore, OCR involves two processes: 1.feature extraction 2.classification. The process of character recognition becomes very tough in the case of Indian languages like Tamil. Many inter-class dependencies exist in Tamil. In Tamil language, many letters look alike. So classification becomes a big challenge.

In image processing, Hough transform are used for feature extraction. Also in recent years, statistical method has gained momentum in the field of classification, which are very essential in OCR. In addition, the combination of statistical method and Hough transform has given good results for font recognition and facial recognition. In this paper, a method is proposed for character recognition, which uses Hough transform to extract features and statistical method for classification. The proposed method validate training methodology. In training phase, the features extracted using Hough transform is fed to Statistical method

II. OCR PROCESS

Here is a very- basic overview of how an OCR engine processes an image to return text contained it: An image of the document is acquired by the computer.

1) An image of the document is acquired by the computer.
2) The image is submitted as input to an OCR engine.
3) The OCR engine matches portions of the image to shapes it is instructed to recognize.
4) Given logic parameters that the OCR engine has been instructed to use, the OCR engine will make its best guess as to which letter a shape represents.
5) OCR Results are returned as text.

III. TRAINING STEPS

A. Scanning

The training files are made ready for all the fonts of various styles. Then the training files are scanned and saved as bitmap images.

B. Preprocessing

The scanned image may have some skewness and the whole image on 3-d scale cannot be processed. Therefore, the image is normalized and skewness is corrected and is converted to a binary images.

C. Segmentation

The first process on preprocessed image is segmentation where we use the following algorithm to extract all characters from the image. Segmentation was performed in two phases. (i) line segmentation wherein each line in the document was segmented using horizontal profile. (ii) character segmentation wherein each character in the line was segmented using 8-connected component analysis. The two-phase approach was adopted based on comparative study where this approach yielded a better result.

D. Thinning

Skeletonization is the method of detaching off of an outline as a lot of pixels as possible as lack of disturbing the common form of the outline. In other words, subsequent to pixels have been detached off, the outline should still be recognized. Hence we use the method for detaching the outline by the Skeletonization process is used to obtain as thin as possible, connected and centered when these are satisfied, then the algorithm must stop

E. *Feature Extraction*

A popular feature extraction method used in digital image processing is the Hough transform (HT). It is able to detect straight lines, curves, or any particular shape that can be defined by parametric equations. In essence, this method maps the figure points from the picture space to the parameter space and, thereafter, extracts the features. There are two types of Hough transform: standard HT and randomized HT. The sequel briefly expounds these two types of HT. Randomized Hough transform is an enhanced approach for detecting features in binary images. It takes into account the drawbacks of standard Hough transform, namely long computation time and large memory requirements. It is based on the fact that a curve or a shape can be defined in the parameter space with a pair or *n*-tuple of points (depending on the shape to be detected) from the original binary picture. Consider a set of figure points P = {(*x*1, *y*1), . . . , (*xn, yn*)} in the binary picture. Now, let (θ, ρ) be the two parameters of the lines to be detected. A pair of points (pi(x,y), pj(x,y)) is selected randomly from the setP and mapped to the parameter space. The corresponding accumulator bins are then incremented. Unlike standard Hough transform where each figure point is mapped to the parameter space, randomized Hough transform maps a set of figure points that may or may not form a shape; therefore, a considerable decrease in the computation time is noticed. The result of continuing this process will be the appearance of maxima in the accumulator bins of the parameter space. These maxima can thereafter be used to detect the lines in the binary pictures.

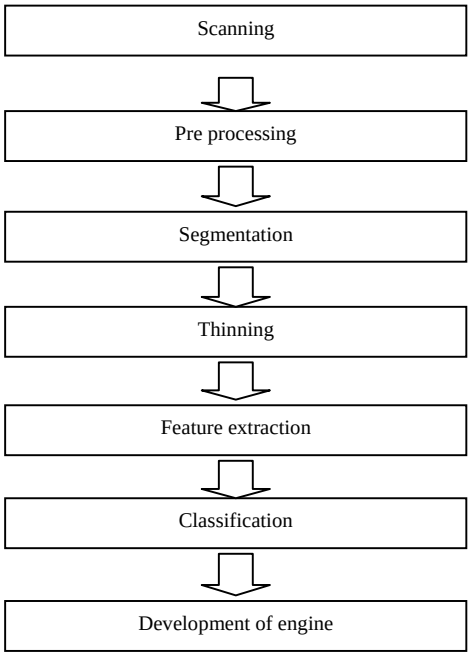

Fig. 1 Flow Chart for Training

F. *Classification Using Statistical Method*

Statistical decision theory is concerned with a set of optimality criteria, and statistical decision functions which maximizes the probability of the experimental pattern given the model of a certain class. Statistical techniques are, mostly, based on three major assumptions:

1) Distribution of the feature set is Gaussian or in the worst case uniform,
2) There are sufficient statistics available for each class,
3) Given ensemble of images {I}, one is able to extract a set of features {fi} ∈ F, i ∈ {1... n}, which represents each distinct class of patterns.

The measurements taken from n-features of each word unit can be idea to represent an n-dimensional vector space and the vector, whose coordinates correspond to the dimensions taken, symbolize the original word unit. The major statistical approaches, applied in the Character Recognition field are the followings:

1) Non-parametric Recognition

The finest known method of non-parametric categorization is the Nearest Neighbor (NN) and is widely used in Character Recognition. An incoming pattern is classified using the cluster, whose center is the minimum distance from the pattern over all the clusters. It does not involve a priori information about the data.

2) Parametric Recognition

Since a priori data or information is available about the characters in the training data, it is possible to obtain a parametric model for each character. Once the consideration of the model, which is based on some probabilities, is obtained, the characters are classify according to some decision rules such as Baye′s method or maximum likelihood.

G. *Development Of Engine*

Once the recognition process is over, the statistical method is ready to classify any kind of new input. The engine can be developed by implementing the statistical method using any programming language.

IV. CONCLUSION

This paper presented an efficient algorithm for classification of characters using Hough transform and statistical method. The system was applied for the recognition of scanned images for cursive English and Tamil language characters, which included the entire character set. The method works brilliantly for frequently used cursive English script and Tamil script. The algorithm did prove more efficient and can be a suitable alternative for the optical character recognition when compared to existing systems.

REFERENCES

[1] J. Cai and Z-Q Liu, "Integration Of Structural and Statistical Information For Unconstrained Handwritten Numeral Recognition," Ieee Trans. On Pattern Anal. Mach. Intell., Vol. 21, No. 3, Pp. 263-270, Mar. 1999.

[2] V. Bansal and R.M.K. Sinha, "On How To Describe Shapes Of Devanagari Characters And Use Them For Recognition", Proc. 5th Int. Conf. Document Analysis And Recognition, Bangalore, India, Pp. 410-413, Sept. 1999.

[3] P. Chinnuswamy, S.G. Khrishnamoorthy, "Recognition Of Handprinted Tamil Characters", Pattern Recognition, Vol. 12, Pp. 141-152, 1980.

[4] P. A. Devijier, J. Kittler, Pattern Recognition,: A Statistical Approach, Prentice Hall, 1982.

[5] Nafiz Arica, T. Yarman-Vural, "An Overview Of Character Recognition Focused On Off-Line Handwritting".[28] A. K. Jain, R. P. W. Duin, J. Mao "Statistical Pattern Recognition: A Review", Ieee Trans. Pattern Analysis And Machine Intelligence, Vol 22, No 1, Pp. 4-38, 2000.

[6] Bharath A, S Madhvanath, Hidden Markov Models For Online Handwritten Tamil Word Recognition. Proc. Intl. Conf. Doc. Anal. Recog. Vol 1, Pp 506-510, 2007.

[7] Hp Labs Isolated Handwritten Tamil Character Dataset. Http://Www.Hpl.Hp.Com/India/Research/Penhwinterfaces-1linguistics.Html#Datasets 1220

[8] Rafael M. O. Cruz, George D. C. Cavalcanti And Tsang Ing Ren "An Ensemble Classifier For Offline Cursive Character Recognition Using Multiple Feature Extraction Techniques" Ieee 2010.

[9] Hough Transform Based Fuzzy Feature Extraction For Bengali Script Recognition Pattnern Recognition And Pattern Recognition Lett. 20 (1999) 771-782.

[10] G. Siromony, R. Chandrasekaran, M. Chandrasekaran, Computer Recognition Of Printed Tamil Characters, Pattern Recognition 10 (1978) 243-247.

[11] Quing Chen, Evaluation Of Ocr Algorithms For Images With Different Spatial Resolution And Noises, Master Thesis, School Of Information Technology And Engineering, University Of Ottawa, 2003.

[12] V. Vapnik, "The Nature Of Statistical Learning Theory", Springer Verlang, 1995.

[13] Rafael M. O. Cruz, George D. C. Cavalcanti And Tsang Ing Ren "An Ensemble Classifier For Offline Cursive Character Recognition Using Multiple Feature Extraction Techniques" Ieee 2010.