ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

A Review on Feature Extraction Techniques for Optical Character Recognition

Neha J. Pithadia1, Dr. Vishal D. Nimavat 2

M.E., Research Scholar, Department of Electronics and Communication Engineering, V.V.P. Engineering College, Rajkot, Gujarat, India[1]

Associate Professor, Department of Electronics and Communication Engineering, V.V.P. Engineering College, Rajkot, Gujarat, India[2]

ABSTRACT: Wide range of applications and numerous other complexities involved in character recognition makes it a continuous and open area of research. Optical Character Recognition is a very important task in Pattern Recognition.Recently Indian Handwritten character recognition is getting much more interest and researchers are contributing a lot in this field. Selection of a featureextraction method is very important factor for high recognition performance in character recognition systems. As different feature extraction techniques are designed for different representation of characters. As an important component of pattern recognition, feature extraction has been achieved close attention by many scholars, and currently has become one of the research spots. This paper gives a general discussion of feature extraction techniques used in optical character recognition.

KEYWORDS: Character Recognition, Feature Extraction, OCR

I. INTRODUCTION

Humans recognize characters easily and they repeat the character recognition process thousands of times every day as they read papers and books. However, after many years of intensive investigation and research, the main goal of developing an optical character recognition system with the same reading capabilities as humans still remains unachieved. Character recognition is the most challenging research areas in image processing. Nowadays different methods are in widespread use for character recognition. OCR will improve the communication interface between man and machine. It is able to convert machine printed or hand written document into editable text format. Major Steps in an OCR System are described in fig1.1.

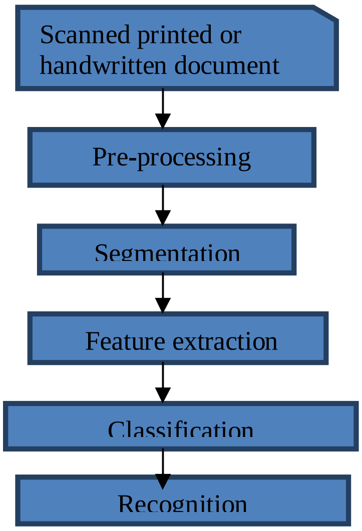

Fig.1 OCR System

Optical character recognition techniques mainly divided into two main parts.

A. *On-line recognition*

Online handwriting recognition has achieved interest due to increase in usage of hand held devices. Those all techniques use all points per stroke for calculating the similarity measurement of characters. The incorporation of keyboard being difficult for the hand held devices demands for alternatives, and in this respect, here in the online method of giving input with stylus is being popular. On-line character recognition more information is available. A few research and studies on on-line and off-line data gives superior recognition performance for on-line data.

B. *Off-line recognition*

Off-line recognition operates on pictures generated by scanner. In the offline recognition data is two-dimensional and space ordered which means that overlapping characters cannot be separated easily and effectively. Off-line character recognition involves the automatic conversion of image into editable text format. Off-line handwriting recognition is very difficult, as people have different handwriting styles. And now a days, OCR machines are primarily focused on machine printed texts.

II. Related work

In paper [1], proposed a review of the character recognition techniques available. Here in this paper it is shown that the character recognition is mainly divided into main two headlines online character recognition techniques and offline character recognition technique. Than the online character recognition is further divided in to two categories 1) K-NN classifier 2) Direction based. Than the offline character recognition is divided in the four categories 1) Clustering 2) Feature Extraction 3) Pattern matching 4) Artificial neural network. This method is again classified further. All the methods are listed is discussed throughout the paper. In paper [2],it proposed a review on the techniques available for the optical character recognition here mainly it is divided under the two main headlines one is online character recognition and another is offline character recognition. Than in the offline character recognition technique there is division like the magnetic character recognition and optical character recognition again in the optical character recognition there can be a handwritten or a printed documents for the recognition. Here word optical character recognition is used because for the process of character recognition the document is firstly scanned. This paper provides useful guidance for the readers working in the area of character recognition. In paper [3], both offline and the online cases have been considered. Here this paper provides the information about the handwritten character recognition. In this paper the algorithms for the pre-processing, word and character recognition and performance with the practical systems are indicated application of character recognition is like signature verification, writer authentication. This paper provides useful guidance for the pre-processing steps involved in the process of character recognition.

III. FEATURE EXTRACTION

The main idea of the feature extraction techniques is to identify characters based on features that are similar to the features humans use to identify characters. Developers or programmers have to manually determine the properties of characters they feel are important. Some properties as example Aspect Ratio, pixels above horizontal half point, pixels to right of vertical half point, distance from image centre, reflected y axis , Is reflected x axis. Researchers have used many techniques of feature extraction for handwritten characters. Shadow code, fractal code, profiles, moment, template, structural, wavelet, directional feature etc., have been addressed in the literature as features. Selection of a feature extraction technique is probably the single most important factor for high recognition performance in character recognition systems. Feature extraction techniques are classified into three major groups as [4].

Statistical features.
Global transformation and series expansion
Geometric and topological features

Statistical feature

1. *Projection Method*

In the projection method it compares data through a projection. Black pixel counts are taken along parallel lines through the image area to generate distributions. The direction of projection will be horizontal axis, vertical axis, diagonal axis or all of the above. The character can be divided vertically and horizontally into four parts and do the same projection on each quarter. From that it will improve the recognition rate[1]. Fig. 2 shows horizontal and vertical projection.

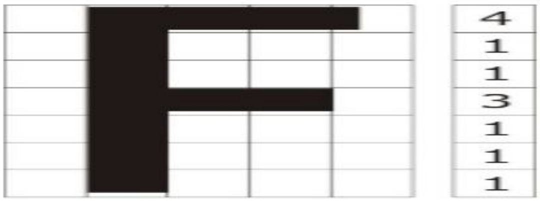

Fig. 2 Projection method

2. *Border Transition Technique (BTT)*

In border transition technique it assumes that all the characters are oriented vertically. Each character is divided into four equal quadrants. The scanning and calculation of zero-to-one transition in both vertical and horizontal directions in each division take place.

3. *Zoning*

Zoning is a method involves the division of the character into smaller fragment of areas. The black pixels in each zone are counted and accumulating or averaging the profiles in each zone extracts features of character[1]. Fig. 4 shows the zoning.

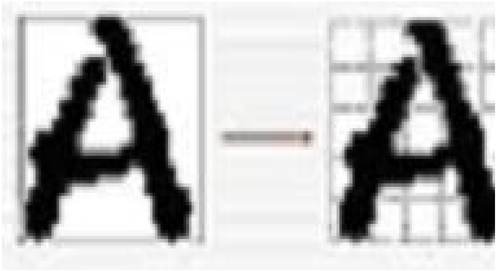

Fig. 3 zoning

4. *Graph Matching Method*

In a graph matching method it uses structural feature of character. It is useful method to change of font or rotation. In this three features are defined. Here first, an end point is connected only one pixel which has information of position. Than a branch point is connected more than three pixels having feature information which is connected the branch point. The information includes of features like, position and direction A curve point is connected two pixels. However a straight line is also connected two pixels. In order to distinguish between a curve point and a straight line, direction information is used.

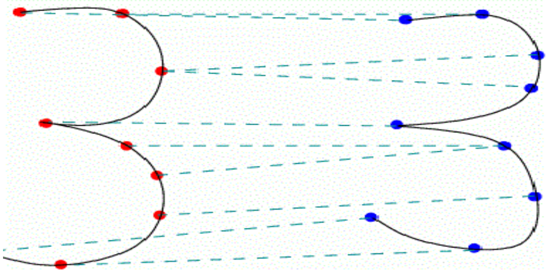

Fig. 4 Graph matching method

Global transformation and series expansion

In this various techniques are:
1) Fourier transforms
2) Gabor transforms
3) Fourier Descriptor
4) Wavelets
5) Moments
6) Karhunen-Loeve expansion etc.

1) Fourier Descriptors

Shape feature vector consists of the Fourier descriptors. After boundary the pixel set of an object was computed. In fourier descriptor technique it uses centroid distance to decide shape signature from boundary. Here this centroid distance function is the periodic function we consider and decompose into fourier series. Fourier transform is used for shape to determine Fourier coefficients, and pixel brightness will find out into computational process of the Fourier coefficients so that shape features can be computed. The Fourier coefficients that we have find out are invariant to translation, scaling, rotation and change of start point are used as Fourier descriptors [14].

Note: Fourier series means decomposing a periodic function into sum of set of sine and cosine functions. The coefficients corresponding to are the Fourier coefficients.

• If than we want to compute invariance to translation, not to use DC term, that is the first element in your resulting array of Fourier coefficients f[0].

• Than if we want to compute invariance to scaling, make the comparison ratio-like, for example by dividing every Fourier coefficient by the DC-coefficient. f*[1]= f[1]/f[0] and so on.

• Than if we want invariance to the start point of contour, only use absolute values of the resulting Fourier coefficients. From that shape features are extracted. The method may also use complex coordinates and curvature function as shape signature.

2) Wavelets

In the wavelets Transform it represents a mathematical way used to study non-stationary signals. Therefore, its usefulness has been increasingly adapted over the last 10 years. It was employed in different fields such as communication technology, geophysics and image processing. The wavelet transform provides an appropriate basis for image handling because of its useful features. The assets of the wavelet transform are:

The ability to compact most of the signal’s energy into a few transformation coefficients, which is called “energy compaction” and the ability to capture and represent effectively low frequency components (image Backgrounds) as well as high frequency transients (image edges). Wavelet transform coefficients are energy, variance and length of waveform. The features are extracted from these coefficients. For the commonly used discrete Fourier transform(DFT) use of DWT feature extraction is an alternative option. Feature vectors belonging to separate signal segments are then classified by a competitive neural network as one of the methods of analysis and processing. By help of wavelet analysis, a matrix of data is obtained, where time and frequency domain information is present. Another waveform is compressed or stretched to obtain wavelets of different scales that are used along time comparing them with the original signal [15].

3) Moments

This technique named moment based features is very effective in describing shape of characters. It was observed that this features can become very effective if some operations such as normalization of size of character and geometric operations are performed correctly by floating point arithmetic. And we use the features drawn by invariants moment technique which is used to evaluate seven different parameter of a character. Moment invariants are known as to be invariant under translation, scaling, rotation, reflection. These are the measures of the pixel distribution around the center of gravity of the character. Among the several moment families introduced in the past, the orthogonal moments are the most popular moments widely used in many applications, owing to their orthogonally property[17].

Geometric and topological features

This method extracts the geometric features of the character. Those features are based on the basic line types that form the character skeletons. This system gives a feature vector as its out- put. The various steps involved in geometric method are:

i) Initially in preprocessing (binarization, skeletonization) is done on the input image.
ii) Than universe of discourse will be selected the features extracted from the character image will include the positions of different line segments in the character images.
iii) After the step (ii) the image will be divided into equal size of windows, and the feature is done on individual windows.
iv)To extracting the different segments of line in a particular zone, the entire skeleton in that zone will be traversed. So that fixed pixels in the character skeleton were defined as starters, intersections.
v) After that the line of each segment is determined, based on this information feature vector is formed and each of the zone has a feature vector corresponding to it. The con- tents of each zone feature vector are

- No. of horizontal lines
- No. of vertical lines
- No. of Right diagonal lines
- No. of Left diagonal lines
- Normalized Length of all horizontal lines.
- Normalized Length of all vertical lines.
- Normalized Length of right all diagonal lines

IV. Conclusion

In this paper the feature extraction methods which are used for other Indian languages and can be used are discussed. This paper represents a review of the feature extraction techniques available for the optical character recognition. As the efficiency is based on the technique we use for the extraction of the features all the techniques will give different attributes of the character.

References

1. Suruchi G. Dedgaonkar, Anjali A. Chandavale, Ashok M. Sapkal , “Survey of Methods for Character Recognition”,International Journal of Engineering and Innovative Technology (IJEIT) Volume 1, Issue 5, May 2012.
2. Er. Neetu Bhatia, “Optical Character Recognition Techniques: A Review”International Journal of Advanced Research in Computer Science and Software Engineering ,Volume 4, Issue 5, May 2014 ISSN: 2277 128X.
3. ReÂjeanPlamondon, “On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey” Fellow, IEEE, and Sargur N. Srihari, Fellow, IEEE, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, Vol. 22, no. 1, January 2000
4. AshlinDeepa R.N, R.RajeswaraRao,“Feature Extraction Techniques for Recognition of Malayalam Handwritten Characters: Review”International Journal of Advanced Trends in Computer Science and Engineering, Vol. 3, No.1, Pages: 481– 485 (2014)
5. Nisha Sharma, TusharPatnaik, Bhupendra Kumar, “Recognition for Handwritten English Letters: A Review”,International Journal ofEngineering and Innovative Technology (IJEIT) Volume 2, Issue 7,January 2013.
6. Ivind Due Trier, Anil K. Jain, and TorfinnTaxt, “Feature Extraction methods for character recognition: A survey”, Department ofComputer Science, Michigan State University, A714 Wells Hall, EastLansing, MI 48824{1027, USA Revised July 19}, 1995.

7. Pritpal Singh, SumitBudhiraja, " Feature Extraction and Classification Techniques in O.C.R.Systems for Handwritten Gurumukhi Script – A Survey", International Journal of Engineering Research andApplications, (IJERA) ISSN: 2248-9622, Available: www.ijera.com

8. Vol. 1, Issue 4, pp. 1736-1739

9. Kartar Singh Siddharth , Mahesh Jangid, RenuDhir, Rajneesh Rani, "Handwritten Gurmukhi Character Recognition Using Statistical andBackground Directional Distribution Features", Innovative Systems Design and Engineering, ISSN 2222-2871 (Online) Vol 3, No 3, 2012.

10. Satish Kumar, "Neighborhood Pixels Weights-A New Feature Extractor", International Journal of Computer Theory and Engineering,Vol. 2, No. 1 February, 2010 1793-8201

11. M. H. Glauberman, "Character Recognition for Business Machines", Electronics 29, 1996, pp.132-136

12. K. M. Kim, J.J. Park, Y.G. Song, I. C. Kim and C. Y. Suen, "Recognition of Handwritten Numerals Using a Combined Classifier with Hybrid Features", SSPR & SPR, LNCS 3138, 2004, pp. 992-1000.

13. N. Arica and F. T. Yarman-Vural, "Optical Character Recognition for Cursive Handwriting", IEEE Transactions on Pattern Analysis andMachine Intelligence,2002, vol. 24, no. 6.

14. M. Bokser, \Omni document technologies," Proceedings of the IEEE, vol. 80, pp. 1066-1078, July 1992

15. Gang Zhang ; Coll. of Inf. Sci. & Eng., Northeastern Univ., Shenyang ; Ma, Z.M. ; Qiang Tong ; Ying, " The shapefeature extraction usingfourierdesciptors with Brightness", presented at the, IIHMSP '08 International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 2008, Page(s):71 - 74 Print ISBN:978-0-7695-3278-3

16. Kheder G., Kachouri A., Taleb R., Ben Messaoud M. and Samet M., "Feature extraction by wavelet transforms to analyze the heart ratevariability during two meditation technique", presented at the sixth WSEAS International Conference on Circuits, Systems, Electronics, Control & Signal processing, Cairo, Egypt,Dec 29-31, 2007.

17. R. J. Ramteke, "Invariant Moments Based Feature Extraction for Handwritten Devanagari Vowels Recognition", 2010 InternationalJournal of Computer Applications (0975 - 8887) Volume 1 – No. 18.

18. G.A. Papakostas, D.E. Koulouriotis and V.D. "Tourassis Feature Extraction Based on Wavelet Moments and Moment Invariants in Machine Vision"

19. Dinesh Dileep, "A Feature Extraction Technique Based on Character Geometry for Character Recognition", Arxiv, 2012.