DOI: http://dx.doi.org/10.26483/ijarcs.v9i2.5868

ISSN No. 0976-5697

Volume 9, No. 2, March-April 2018

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

FEATURE EXTRACTION TECHNIQUES FOR HANDWRITTEN CHARACTER RECOGNITION

Madhuri Yadav
USCI&T
GGSIPU
New Delhi, India

Alok Kumar
Department of computer science
CDAC
Noida, India

Abstract: In this era of growing automation, automatic character recognition has become need of the hour. Optical character recognition (OCR) techniques allow computers to recognize handwritten characters and change them in digital format or other formats which are understandable by computers. This automatic recognition requires feature extraction from character images. Feature extraction is a very important phase of character recognition. This paper discusses features based on shapes, skeletons, image moments, image transforms, critical points, etc. This paper also investigates different types of features used in literature works. A good feature set results in good recognition rates. Thus, it is important to have knowledge of different types of features and their properties. This paper benefits its readers by providing them insight of different types of features and helps them in identifying the appropriate feature set according to their application.

Keywords: OCR; feature extraction; Handwritten characters; Offline recognition; character recognition; classification

I. INTRODUCTION

Automatic character recognition is conversion of handwritten documents in computer editable format. Character recognition has wide number of applications in various domains including banking sector, educational institutions, postal sector, and other industries with high paper work. In banking, automatic recognition of numerals and letters written on cheques and other documents can highly elevate the work capacity. Similarly, in post offices, universities and colleges automatic form processing, and checking can reduce paper load and thus save environment. Apart from these benefits, automatic recognition system also helps in restoring old degraded books. It is also an aid for physically challenged person, as it can be used for synthetic speech generation. There are two types of character recognition techniques namely, offline and online recognition techniques. Offline recognition techniques work on static images, whereas, online recognition systems work with the characters collected through digital equipments such as touchpads, PDAs, tablets etc. Thus, online recognition systems work on real time information and extract features more accurately and easily than offline systems. The main objective of this paper is to investigate the features used for offline recognition. Fig. 1 highlights the basic phases of character recognition which are require prior to feature extraction.

A. Data Collection

Data collection is a process of acquisition of data from different sources. For character recognition it can be collected from forms, cheques, postal envelopes, or other documents. The collected data must contain characters written by people of different age groups, sex, education and time. The different languages have benchmark databases for standardization of entire recognition procedure.

ISI Kolkata database [1] for hindi handwritten characters was collected from mail, job application forms, specially designed forms, and cheques. The most widely used database for arabic language is IFN/ENIT database [2] which contains city names. Chinese handwritten character recognition are mainly based on CASIA-HWDB1.0 and 1.1 [3].

B. Pre-processing

This phase is used to enhance the quality of images. Pre-processing phase also plays a vital role in recognition rates because noise free data will result in good feature extraction which in turn gives better classification results. It includes noise reduction, smoothing, sharpening, thinning, skeletonization, normalization, skew detection and correction, contour analysis, binarization etc.

The purpose of feature extraction is to prepare images with optimal quality so that feature extraction and classification can work efficiently and correctly.

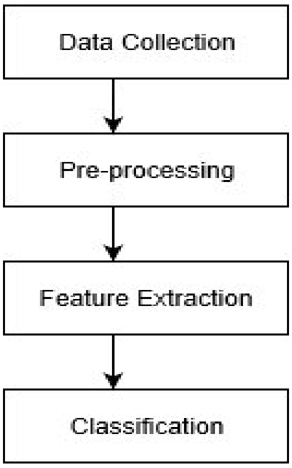


Figure 1. Basic phases of character recognition

C. Feature Extraction

Features are the most pertinent set of attributes which define character images. Instead of using entire set of pixels of images, only a few critical pixels are extracted from images and used as features. This phase is discussed in detail in further sections of this paper.

D. *Classification*

The main aim of character recognition is assigning class labels with character images. There are various classification methods which can be classified as statistical, structural, supervised and unsupervised. Statistical methods include Bayesian classifiers, Guassian classifiers, Quadratic discriminant function(QDF) and Modified Quadratic discriminant function(MQDF). Statistical classifiers include decision trees, graph matching, string matching, expression generation etc. Support vector machines (SVM), multilayer perceptrons are based on supervised learning. k-nearest neighbor (k-NN), autoencoders are unsupervised classifiers. Nowadays, combination of these classifiers are also used in various applications.

II. FEATURE EXTRACTION

This section describes popular feature extraction techniques which have been used in various research works. The features extraction methods can be broadly classified as statistical, structural, image transforms as shown in Figure 2. A good feature set facilitates classification step. It is also a type of dimensionality reduction which efficiently represents critical parts of the image and hides irrelevant details of image. Feature extraction can be local or global depending upon the part of the image used for feature extraction. Local feature extraction involves zoning of image i.e. dividing image into smaller sections according to some topology or fixed zones. There can be different types of zones as discussed in [4]. Global feature extraction extracts features from complete image. The features such as geometric moments have significant results when extracted from complete image. There are some global features which can be applied locally, for example, calculation of centroid can be done for entire image as well as for different zones of image.

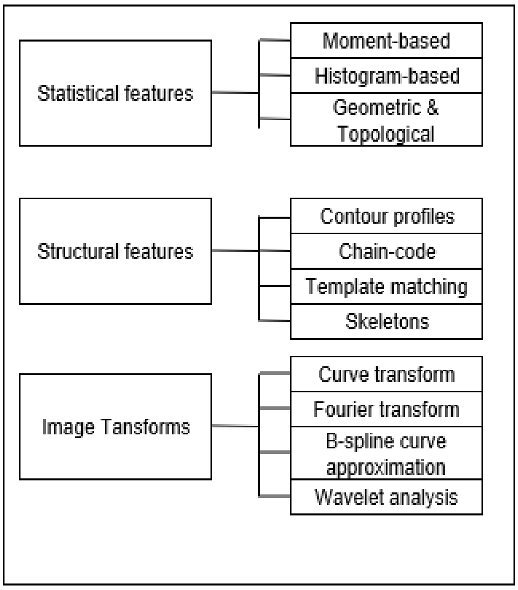


Figure 2. Broad classification of feature extraction technique

A. *Statistical Features*

These features include invariant moments such as Hu-geometric moments, zernike moments, legendre moments. Histograms of directions, gradients, angles, oriented gradients are also important features of this class. Geometric and topological features involve extraction of critical points, lines or arcs from image, such as end points, branch points, loops, diagonal lines, extreme points etc.

Moments are scalar quantities used to characterize a function and to capture its significant features. They have been widely used in almost all domains of pattern recognition. Hu [5] gave seven geometric moments which were invariant to scaling, translation and rotation. The magnitudes of a set of orthogonal complex moments of the image are known as Zernike moments [6]. Legendre polynomials form the basis function in legendre moments. Legendre moment for (NxN) image is given by (1).

$$L_{pq} = \sum_{m=0}^{N-1}\sum_{n=0}^{N-1} P_p(m_N)P_q(n_N)\, f(m,n) \qquad (1)$$

Where, $P_p(x) = \frac{1}{2^p\, p!}\frac{d^p}{dx^p}\left(x^2-1\right)^p, x \in [-1,1]$ and $m_N = \frac{2m-N+1}{N-1}$ and p and q are integers between (0,infinity).

Histograms are charts displaying the distribution of a set of pixels according to some criteria like gray scale values, direction of pixels, gradient direction etc. Different numbers of bins are created for oriented gradients in Histogram of oriented gradients. Depending upon the application different types of histogram bins can be generated and used as features.

Topological features explore topology of character and extracts critical points from the character image. It is easy to use these features once the input image is transformed in skeleton or contour traced. The critical points include end points, branch points and cross points. They are extracted from binary image or black-white image. Encircled areas in Fig. 3 illustrate the three feature points for hindi handwritten character. The end points are the extreme points of line segments. The branch point is a junction point containing three branches and the last point i.e. cross point is junction point containing four branches.

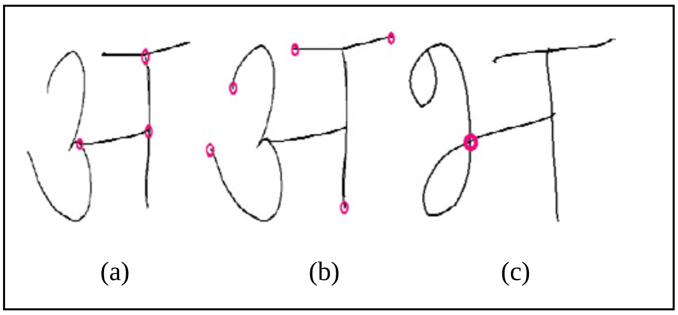


Figure 3. Types of points: (a) Branch points; (b) End points; (c) Junction point.

B. *Structural features*

These features are generally global features. They include contour tracing or border following, chain code detection, template matching, regular expression, graph matching etc. Contour tracing extracts information about general shape of a character. It can itself be used as feature or it can also be used as pre-processing step for extracting features. It extracts boundary of a character giving a small subset of pixels which represent complete character. The two most famous tracing algorithms are square tracing algorithm and Moore-neighborhood tracing algorithm. Fig. 4 illustrates boundary tracing using octagon. Similarly, chain code and skeletonization can be used as features. Freeman chain codes [7] can be 4 or 8-directional as shown in Fig. 5. They represent direction of pixels within local neighborhood. Differential chain codes, a variant of freeman chain code are more promising as they are rotation invariant. Contours or skeleton can be polygon approximated into piecewise line segments or curve segments. Since, contours contains large number of pixels, so polygon approximation can be used to describe essential contour shape. Template matching can be used to compare regular expressions generated using skeletons. Speed up robust features (SURF) and Scale invariant feature transform (SIFT) can be used with template matching.

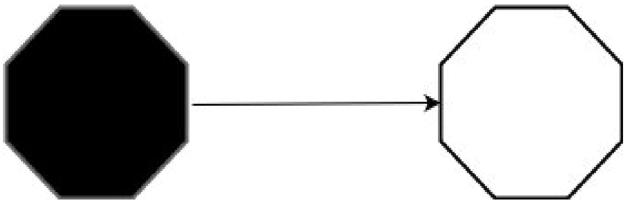
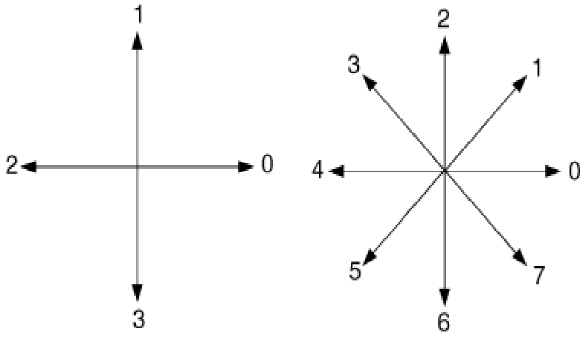


Figure 4. Boundary tracing using moore-neighborhood tracing algorithm.

Figure 5.



(a) (b)

Figure 6. Freeman chain codes: (a) 4-directional; (b) 8-directional.

C. *Image Transforms*

Popular feature extraction methods used in optical character recognition are image transformations. The above discussed methods used spatial domain for feature extraction whereas, image transformations such as fourier descriptors used frequency domain. Image transforms include Hough transform (HT), curve approximation, and wavelet transform.

Hough transform [8] detect straight lines, curves, or any particular shape that can be defined by parametric equations. Fourier descriptors (FD) is one of the most popular and efficient method and it represents the shape of the character in the frequency domain. B-spline curves and curve transform can be obtained by cubic or polynomial approximation of data points. Curves are controlled by data points (which are subset of character points)also called control points. These control points determine the slope of segments and determine the overall shape of the character.

Wavelet transform uses functions that are localized in both real and fourier space. Discrete wavelet transform(DWT) returns a data vector of same length as the input is. DWT decomposes image into wavelets (functions) that are orthogonal to translation and scaling.

III. Conclusion

This paper gives brief introduction about phases of character recognition. It discusses feature extraction phase and gives detailed overview of type of features used for character recognition. Categories of the discussed features are statistical, structural and image transforms. All feature extraction techniques have their potential and limitations, so they are to used according to application.

IV. References

[1] U. Bhattacharya and B. B. Chaudhuri: “Databases for research on recognition of handwritten characters of indian scripts”, International Conference on Document Analysis and Recognition(ICDAR’05), Washington, DC, USA, Aug 2005, pp. 789–793.

[2] M. Pechwitz, S. S. Maddouri, V. Margner, N. Ellouze, H.Amiri, “IFN/ENIT database of handwritten Arabic words”, In 7th Colloque International Francophone sur lEcrit et le Document, 2002, pp. 129–136.

[3] C. Liu, F. Yin, D.Wang, Q.Wang, “CASIA online and offline Chinese handwriting databases”, Proceedings of International Conference on Document Analysis and Recognition (ICDAR), Beijing, China Sept 2011, pp. 37–41

[4] D. Impedovo and G. Pirlo, “ Zoning methods for handwritten character recognition”, Pattern Recogn. 47, 2014, 969-981.

[5] M.K.Hu, “Visual pattern recognition by moment invariants,” IRE Trans. Inf. Theory , 1962, pp. 179–187.

[6] A. Khotanzad and Y. H. Hong, “Invariant image recognition by Zernike moments”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12(5), 1990, pp.489–490.

[7] H. Freeman. Boundary encoding and processing. In B. Lipkin and A. Rosenfeld, editors,Picture Processing and Pshcholopictorics. Academic Press, New York, 1970, pp. 241–266.

[8] Kultanen, L. Xu, and E. Oja. Randomized Hough transform. In Proceedings of the 10th International Conference on Pattern Recognition, Atlantic City, vol. 1, 1990, pp. 631–635.

© 2015-19, IJARCS All Rights Reserved