Handwritten Devanagari Compound Character Recognition Using Zernike Moment

[1]Juhee Sachdeva,[2]Dr. Shilpa Sharma
[1]Ph.D.Scholar, Jaipur National University,Jaipur,India,[2]AP JNU Jaipur India.

Abstract

Character recognition of Devanagari compound characters is a difficult task because characters are written in half forms of original characters. In this paper, we present OCR for HandwrittenDevanagari compoundcharacter. Recognition of Devanagari compound characters consists of three main phasesPreprocessing,Feature Extraction and Classification. During last few decades many research has been conducted to find a solution for Character recognition as it is a solution for many real world problems. The major problem for processing large volume of data can be solved.Work has been done in recognizing handwritten characters in Languages like English,Chinese,Japanese and Arabic. This paper focuses on the recognition of Devanagari compound characters. The feature extraction stage uses Zernike Moments based feature extraction technique that analyzes and select set of features that can be used to uniquely identify the character. Zernike Moment based techniques uses rotation invariance that is successfully applied to many Pattern Recognition problems and character recognition system. Moment based system uses global features that is suitable for compound character recognition. For classification phase, Support Vector Machine is used . The system reported a high recognition rate of accuracy.

Keywords

Compoundcharacters, Pattern Recognition,Zernike Moments,Support Vector Machine.

INTRODUCTION

Handwritten character recognition converts handwritten images into text that can be stored and processed further. Handwritten character recognition is quiet popular due to its application areas which would reduce the task of data entry and save the time in case of Form filling, Postal Automation, and Banking etc[1]. It can also be used for reading of addresses on envelops and reading of ancient handwritten documents. All the research that has been done so far are for simple characters i.e. consonants and vowels. But Devanagari script consist of one more set of characters that are called Compound characters that are formed by joining two consonants together. Feature of compound characters are more complex as compared to simple characters and hence recognition of these characters is a complex task.Compound characters occur frequently in Devanagari script so we can not ignore this area of character recognition system. Devanagaricharacters has complex shapes and it vary with the mental and physical condition of writer that makes it further more complex problem. Other factors like pen ink color, pen width and acquisition device may affect the shape of a character. Hence handwritten Devanagari characters are more complex due to their shape and structure[3][4]. Devanagari characters have loops, curves and complex structures. Few character sets have similar shapes that made it more complex in recognition. Compound characters recognition is further more complicated as their separation and identification is more difficult[16][17].

Devanagari Script

Devanagari is the most popular script in India. Devanagari character set consist of 11 vowels and 33 consonants. Consonants are the basic character set of the script whereas Vowels can be written as independent letters or can be used as modifiers. Modifiers are symbols that can be used in connection with consonants. These modifiers can be used above, below, before or after to a basic character.Vowels become modifiers as soon they are written in this way and the characters so formed are called conjuncts. Sometimes consonants can be written in their half form and combine with another consonants, these form of characters are called compound characters[5].Devanagari characters can be written and classified on the basis of vertical bar present in a character.

Compound Characters with End Bar Character

स्त ख्य ज्य
त्व म्य च्य
ल्प न्म न्न
त्प ल्य श्ल
घ्य त्य स्व

Compound Characters with Middle Bar Character

क्क क्फ

Compound Characters with No Bar Character

टृ ह्ह छ्ड ङट्र

English character does not have such characteristic and so it can be taken as a distinguishable feature to extract English from these scripts[7]. In Devanagari script all letters in a word are joined through a common horizontal line or header line known as shirorekha. All the letters of a word appear as they are hanging over the header line. Devanagari script has more complex character sets that include compound characters, modifiers, consonants and vowels. These characters have complex structure as compared to English characters. Devanagari Text has threezones i.e. Upper zone, middle zone and lower zone[6][8].
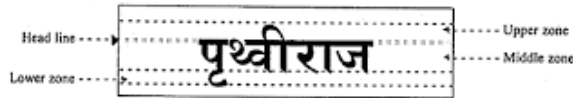

Fig:3 Three zones of a Devanagari word
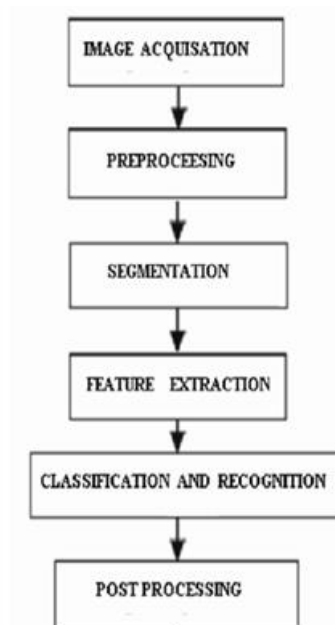
1. Proposed Methodology

Fig:4 Offline Character Recognition System

To recognize the handwritten data from document, there are number of steps which are involved, firstly the document is scanned using scanner[1]. This scanned document is converted into image. Then image is preprocessed with set of valuable steps and convert it into a character/script as per the environment. The scanned image undergoes number of valuable preprocessing steps so as to increase the ratio of recognition of the handwritten document. The general steps for Handwritten Character recognition is Image Acquisition, preprocessing, feature extraction, classification and recognition[3].

1.1 Preprocessing:
Preprocessing consist of various operations performed on image.It enhances the image making it suitable for next level of segmentation.All work has been done in MATLAB[21].

- Image samples are scanned in RGB format.
- RGB image is then converted in gray scale image.Grayscale image is then converted into binary image.

1.2 Segmentation:

In the segmentation phase,an image is decomposed into several sub images i.e. sequence of characters are decomposed in to sub images of individual characters. There are commonly three types of segmentations namely line segmentation, word segmentation and character segmentation.

Line Segmentation: In Devanagari script, horizontal projections is used to segment lines. Each line is identified and segmented. Header line is used to have maximum number of pixels while base line is used to have minimum number of pixels in horizontal projections.

Word Segmentation: After line segmentation, the text in each line is segmented to detect words. It is called,line segmentation, because there is generally enough space presents between words and each word is bounded by header line having no space within word.

Character Segmentation: Now after word segmentation, each word need to segment into characters. It is called character segmentation. In Devanagari script,sheirorekha binds character so we generally use to remove header line first to have vertical space between charactersthe word under consideration. Then segmenting the word into characters is based on present vertical space between these characters.Each individual character is resizes into 100x100 pixels for extracting its features.

1.3 Feature Extraction:

Feature extraction can be considered as finding a set of parameters (features) that define the shape of the underlying character as precisely and uniquely as possible[24]. The term feature selection refers to algorithms that select the best subset of the input feature set. Compound characters are more complex in shape and structure. They consist of variousendpoints,junctions points, curves and strokes. Methods that create new features based on transformations,or combination of original features are called feature extraction algorithms[9][10].In 1961 Hu derived Method of Two-Dimensional Invariants and these moments can be used in character recognition as the classification error rate is quite low. The term 'Invariant' denotes feature which remain unchanged if the image undergoes changes like change in size,position,orientation. We are using Zernike Moment based Feature Extraction technique for proposed work. The Zernike moments magnitude is used as a rotation invariant feature that represent pattern of character. Zernike moments give an orthogonal set of rotation-invariant moments. Teague first introduced the use of Zernike moments that are rotation invariant and can easily constructed to any order to overcome information redundancy due to geometric moments.

The Zernike polynomials are set of complex ,orthogonal polynomials that can handle overlapping of information between the moments.

Zernike polynomials fulfill orthogonally

$$\int_0^{2\pi} \int_0^1 Vnj(r,\theta)Vmk(r,\theta)rdrd\theta = \frac{\pi}{n+1}\delta mn\delta jk$$

Two dimensional Zernike moments:

$$Z_{pq} = \frac{p+1}{\pi}\int_0^{2\pi}\int_0^1 V_{pq}^*(r,\theta)f(r,\theta)rdrd\theta$$

Where p-|q| is even and non negative

And p=0,1,2….,∞ and defines the order and * defines complex conjugate and q is an integer defining angular or rotational dependence.

Zernike moments for low-order

$$Z_{00} = \frac{1}{\pi}m_{00}$$

$$Z_{11} = \frac{2}{\pi}(m_{10} - im_{01})$$

$$Z_{20} = \frac{6}{\pi}(m_{20} + m_{02}) - \frac{3}{\pi}m_{00}$$

Image is mapped to unit disc using polar coordinates. Moments based features are extracted from each zone of character bitmapped image.

1.4   Classification

Classification is the decision making phase of a Handwritten character recognition system that makes use of the features extracted from the previous stage in the process[11][13].In the proposed system Support Vector Machine (SVM) is used for classification. SVM are group of supervised learning methods that can be applied to classification. The standard SVM classifier takes the set of input data classify them in to two distinct classes. For multiclass classification problem, classes are decomposed into multiple binary class problem. Given a training database of M data $\{x_m | m = 1, \ldots, M\}$, the linear SVM classifier is then defined as :

$$f(x) = \sum_j \alpha_j \, x_j * x + b$$

Where $\{x_j\}$ are the set of support vectors. The SVM map input data onto a higher dimensional feature space nonlinearly related to the input space and determine a separating hyper plane with maximum margins between the two classes in feature space. The optimal separating hyperplane can be determined without any computation in higher dimension feature space by using kernel functions. Different type of kernels are used to classify samples of different classes with appropriate margins. Commonly used kernels are

Linear Kernel: $k(x, y) = x.y$

Radial Basis Function: $k(x, y) = \exp(-\|x - y\|^2 / 2\sigma^2)$

Polynomial kernel: $k(x, y) = (x.y + 1)^d$

2.   Conclusion

This paper presents a Recognition system of handwritten Devanagari compound characters. Samples of compound characters are collected from different writers of different age. Zernike Moment feature extraction technique combined with Support Vector Machine proves to be a powerful tools for compound character recognition system. Feature extraction uses multiple features which gives better recognition accuracy. Structural pre classification of characters are doneand then Moment based features are extracted. Feature extraction uses multiple features which gives better recognition accuracy.

The Zernike Moment based feature extraction proposed in the paper has the ability to recognize stimulus patterns if a set of input patterns are repeatedly presented to it, it gradually acquires the ability to recognize these patterns. It is not necessary to give any instructions about the categories to which class the pattern belongs. The implementation of Zernike Moment method with SVM give reasonable results towards recognizing compound characters.

References:

[1] Optical Character Recognition by Rice, Stephen V., Nagy, George, Nartker, Thomas A Springer Science & Business Media, 31-May-1999.

[2] Gonzalez, Rafael C., and Richard E. Woods. "Digital image processing." (2002).

[3] Cheriet, Mohamed, et al. Character recognition systems: a guide for students and practitioners. John Wiley & Sons, 2007.

[4] H.Bunke P.S.P Wang-"Handbook of character recognition and document image analysis." (1997).

[5] Connell, Scott D., R. M. K. Sinha, and Anil K. Jain. "Recognition of unconstrained online Devanagari characters." Pattern Recognition, 2000. Proceedings. 15th International Conference on. Vol. 2. IEEE, 2000.

[6] Bansal, Veena, and R. M. K. Sinha. "On how to describe shapes of Devanagari characters and use them for recognition." Document Analysis and Recognition, 1999. ICDAR'99. Proceedings of the Fifth International Conference on. IEEE, 1999.

[7] Bansal, Veena, and R. M. K. Sinha. "Segmentation of touching and fused Devanagari characters." Pattern recognition 35.4 (2002): 875-893.

[8] Pal, U., and B. B. Chaudhuri. "Indian script character recognition: a survey." pattern Recognition 37.9 (2004): 1887-1899.

[9] Nixon, Mark, Mark S. Nixon, and Alberto S. Aguado. Feature extraction & image processing for computer vision. Academic Press, 2012.

[10] Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. A. (Eds.). (2008). Feature extraction: foundations and applications (Vol. 207). Springer.

[11] Liu, Cheng-Lin, and Hiromichi Fujisawa. "Classification and learning methods for character recognition: Advances and remaining problems." Machine learning in document analysis and recognition. Springer Berlin Heidelberg, 2008. 139-161.

[12] J.Pradeep, E.Srinivasan and S.Himavathi, " Diagonal Based Feature Extraction For Handwritten Alphabets Recognition System Using Neural Network", (IJCSIT), Vol 3, No 1, Feb 2011.

[13] Dinesh Acharya U., N. V. Subbareddy & Krishnamoorthi, " Combined Classifiers In Recognition Of Handwritten Kannada Numerals: A Hybrid Approach", International Journal of Information Technology and Knowledge Management July-December 2009, Volume 2, No. 2, pp. 305-311.

[14] Anita Jindal, Renu Dhir and Rajneesh Rani, " Diagonal Features and SVM Classifier for Handwritten Gurumukhi Character Recognition ", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 5, May 2012 ISSN: 2277 128X .

[15] Hadar I. Avi-Itzhak, Thanh A. Diep, and Harry Garland, "High Accuracy Optical Character Recognition Using Neural Networks with Centroid Dithering ", IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 17, No. 2, February 1995.

[16] Veena Bansal and R.M.K. Sinha, "Integrating Knowledge Sources in Devanagari Text Recognition System", Technical Report, TRCS-97-248, I.I.T. Kanpur, India, 1997.

[17] R. M. K. Sinha, "A Journey from Indian Scripts Processing to Indian Language Processing", IEEE Annals of the History of Computing, pp8-31, Jan–Mar 2009.

[18] Lee, Yuchun. "Handwritten digit recognition using k nearest-neighbor, radial-basis function, and backpropagation neural networks." Neural computation 3.3 (1991): 440-449.

[19] Plamondon, Réjean, and Sargur N. Srihari. "Online and off-line handwriting recognition: a comprehensive survey." IEEE Transactions on pattern analysis and machine intelligence 22.1 (2000): 63-84.

[20] MAJIDA ALI ABED HAMID ALI ABED ALASADI "Simplifying Handwritten Characters Recognition Using a Particle Swarm Optimization Approach" EUROPEAN ACADEMIC RESEARCH, VOL. I, ISSUE 5/ AUGUST 2013 ISSN 2286-4822.

[21] Kauleshwar Prasad, Devvrat C. Nigam, Ashmika Lakhotiya:Character Recognition Using Matlab‟s Neural Network Toolbox,International Journal of u- and e- Service, Science and Technology Vol.6 No. 1, February, 2013.

[22] Anita Pal &Dayashankar Singh, "Handwritten English Character Recognition Using Neural Network"International Journal of Computer Science &CommunicationVol. 1, No. 2, July-December 2010, pp. 141-144.

[23] Gunjan Singh, Sushma Lehri "Recognition of Handwritten Hindi Characters using Backpropagation Neural Network",(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (4) , 2012,4892-4895.

[24] Sandhya Arora, "Combining Multiple Feature Extraction Techniques for Handwritten Devanagari Character Recognition", IEEE Region 10 Colloquium and the Third ICIIS, Kharagpur, INDIA, December 2008.

[25] Yegnanarayana, B. Artificial neural networks. PHI Learning Pvt. Ltd., 2009.