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 6, June 2015

Handwritten Devanagari (Marathi) Compound Character Recognition using Seventh Central Moment

P.E.Ajmire[1], R V Dharaskar[2], V M Thakare[3]

Associate Professor, Dept. of Computer Science, G. S. Science Arts & Comm. College, Khamgaon.(M.S.), India1

Former Director, Disha Education Scociety, Raipur, (C.G), India2

Professor & Head, Dept. Of Comp. Sci., PGTD of CS, S G B Amravati University, Amravati, (M.S.), India3

ABSTRACT: Compound character recognition of Devanagari Script (Marathi language) is one of the challenging tasks since the compound character is combination of one or more characters. These characters can be treated as fusion of two or more characters and hence these are complex in structure. Marathi, Hindi, Sanskrit and Nepali are written with Devanagari script. All these languages have compound characters. The characters may be formed with different sequence of combinations of basic characters such as vowels and consonants. Thus the recognition of compound characters makes this task more challenging to the researcher. So, as compared to English like Roman script, one of the major obstacles in handwritten Marathi character recognition is the large number of complex shaped compound characters. Considering the complexity of the problem, the present algorithm makes an attempt to identify the compound character. This paper discuss the recognition of compound character using combination of 7 Invariant Moment (Rotational Moment) and 7th order Central Moment(Translation Moment). SVM is used as classifier. The overall performance of the proposed system is 93.87%.

KEYWORDS: Compound Character, Feature Extraction, Handwritten character recognition.

I. Introduction

Historically, India is multilingual country and Indian constitution recognizes 22 languages which are written with 11 different scripts and nearly 6000 dialects used in different states of the country. Marathi is one of the widely spoken languages in India especially in the Maharashtra state. Hindi, Sanskrit, Marathi and Nepali use the “Devanagari” script for writing and speaking. During this many compound characters are used. These compound characters make the task of its recognition more difficult.

In Pattern recognition area, Optical Character Recognition is still an active and challenging area of research, specially handwritten Marathi Characters. It is due to the different writing styles, which various from person to person. Compound characters are one of the most frequently used constructs in spoken as well as written Marathi Language. Like in any language, two or more consonants can be combined and pronounced together.
e.g. In English word "fact" consonants "c" and "t" and pronounced together, but in Marathi we do not just write symbols one after other. There is little change done in symbol of first consonant which is to be pronounced half. So, consonant symbol is also written half.

Half symbol(s) + full symbol is called जोडाक्षर (joDAkShar).

A. *Devanagari Characters*

As Devanagari characters[1,12], Marathi characters can be divided into three groups based on the presence and position of a vertical bar, namely:

i. end bar characters

DOI: 10.15680/ijircce.2015.0306078

ii. middle bar characters and
iii. non-bar

The position of the vertical bar is the left most column where number of black pixels is 80% or more of the character height. This is determined by scanning the vertical projection of the image from left to right. The first column where target numbers of pixels are present becomes the position of the vertical bar. Figure 1.shows some end bar characters.

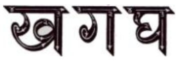

Fig.1 End Bar Characters

In Marathi, there are only three characters written with centred vertical bar shown in figure 2.

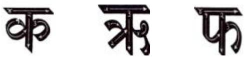

Fig.2 Centered vertical bar Characters.

There are some characters written without any bar are given in figure 3.

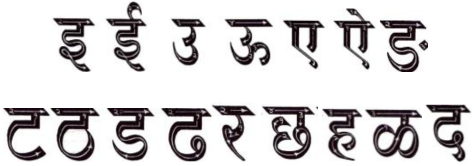

Fig.3 No bar Characters

B. *Compound Character*

In Marathi there is various form of compound character. For each compound character, it is corresponding to its "half" predefined character. For those symbol having vertical in them. Their half-symbol is created by omitting vertical line. The following table shows the half symbol with a 'व' and the corresponding compound character.

Full Symbol	Half symbol	Combining व (v) to half symbol
ख(kh)	ख्	ख्व
ग(g)	ग्	ग्व
घ(gh)	घ्	घ्व
च(ch)	च्	च्व
ज(j)	ज्	ज्व
झ(jh)	झ्	झ्व
ण(N)	ण्	ण्व
त(t)	त्	त्व
थ(th)	थ्	थ्व
ब(b)	ब्	ब्व
भ(bh)	भ्	भ्व
म(m)	म्	म्व
य(y)	य्	य्व
ल(l)	ल्	ल्व

व(v)	व्	व्व
श(sh)	श् or श्‍	श्व or श्व
ष(Sh)	ष्	ष्व
स(s)	स्	स्व
क्ष(kSh)	क्ष्	क्ष्व
ज्ञ(dny)	ज्ञ्	ज्ञ्व

Table 1: Shows Some Compound Characters

In the above table, some compound characters formed by the combination of two basic characters, out of it one is basic character and other is व (v).

Some other form the compound character is shown below.

1. Small cross line connected to the central vertical line of first half character.
   E.g क + र = क्‍र but it is क्र.

2. Small cross line connected to vertical line of the first character.
   E.g प + र = प्‍र but it is प्र.

3. If the first character does not have any vertical then the compound character will have another form.
   E.g ट + र = ट्‌र but it is ट्र.

4. If the first character is र then we have another fom.
   E.g र + ह = रह but it is ऱ्ह

In this way there are 36x36=1296 compound characters form with consonants are available in Marathi. In addition to this there are compound characters formed with vowels i.e. upper and lower modifiers are in Marathi, which are commonly used in the language in written and in spoken language.

II. RELATED WORK

There are many script and Languages in the world. The researchers have done work on some of them like English, Chinese, Latin, Arabic, Japanese, Thai and Devanagari. India is a multi-lingual and multi-script country comprising of eleven different scripts and not much work has been done towards handwritten recognition of Indian scripts. Recognition of handwritten characters is important because of its applicability to a number of problems, like postal code recognition and information extraction from fields of different forms. In the Indian context, there exists a need for development and/or evaluation of the existing techniques for recognition of handwritten character in Indian scripts. A variety of statistical techniques, model and techniques has emerged, influenced by developments in the field of pattern recognition such as character recognition[2,3,4,6], face recognition, finger print recognition, etc. In this paper, we have mentioned the statistical features extraction techniques for its recognition. Selection of a feature extraction method is probably the single most important factor in achieving high recognition performance in character recognition systems. Different feature extraction methods are designed for different representations of the characters, such as solid binary characters, character contours, skeletons (thinned characters), or gray level sub-images of each individual character[8]. The feature extraction methods are discussed in terms of invariance properties, re-constructability, and expected distortions and variability of the characters. When a few promising feature extraction methods have been identified, they need to be evaluated experimentally to find the best method for the given application. A consonant or vowel[10] following a consonant, sometimes takes a compound orthographic shape, which we call as compound character. Compound characters can be combinations of two consonants as well as a consonant and a vowel. Compounding of three or four characters also exists in the script[2]. This becomes the huge set of characters, which is nearly 1500. The complexity of a handwritten character recognition system increases mainly because of various writing styles of different individuals. This is another challenging aspect in the development of recognition system. There are two different ways are proposed for recognition of compound characters [3]. One way by separation of the character and another is without separation of character. i.e. the compound character is consider as a one character.

A. *Database Preparation*

To the best of our knowledge standard dataset of Handwritten Marathi character is not available. Therefore, dataset of handwritten Marathi Characters is design by collecting the handwritten documents from writers of different age groups of different mother tong. There are variations in writing styles. Some writer use ‘Shirorekha’ some doesn’t use ‘Shirorekha’. These contain some basic as well as compound characters with and without Shirorekha. Some words and characters are given in figure 4.

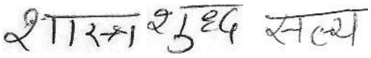

Fig 4. Handwritten words with compound characters

B. *Pre-processing*

Pre-processing is the first step in any character recognition system because this affects the results of method. Pre-processing of compound character involves following steps:

- RGB to Gray Conversion: The handwritten data sheet is scan with resolution 300 dpi in color format of JPG (JPEG) file. For character recognition we need Binary image. So the RGB is converted to Gray and then to Binary.
- Thresholding: The scan image of the character (Gray or Color) is taken as an input to convert into binary image, based on the threshold level[4].
- Binarization: Binarization is an important image processing step in which the pixel values are separated into two groups; white as background and black as foreground. Only two colors, white and black, can be present in a binary. The aim of binarization is to minimize the unwanted information present in the image while protecting the useful information [5].
- Noise Reduction : In image processing, noise reduction and restoration of image is expected to improve the qualitative inspection of an image and the performance criteria of quantitative image analysis techniques Digital image is inclined to a variety of noise which affects the quality of image. The main purpose of de-noising the image is to restore the detail of original image as much as possible [6].
- Normalization: Normalization is An important operation that helps improve the contrast in an image is called Auto-Normalization. The goal of auto-normalization is to improve the contrast of the image by “stretching” the range of intensity values to span a greater range of (luminance) values. Auto-normalization is performed on every image we process. It’s a linear, lossless operation that can take a low-contrast image and make it more legible. On images that have good contrast to begin with, auto-normalization has little to no effect.
- Thinning: Thinning of the character reduced the thickness of the character to one pixel. This process is also called skeletonization. We performed morphological operation ‘skel’ on binary image. [8]

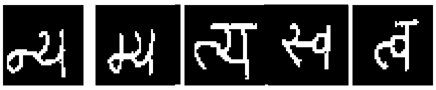

Fig.4 Sample compound characters after pre-processing

III. SEGMENTATION

Segmentation refers to the process in which an image is subdivided into constituent regions or objects. These objects can be further processed or analyzed for the extraction of quantitative information. There are various types of segmentation which are paragraph segmentation, line segmentation, word segmentation and character segmentation. Character segmentation is a procedure in which from the word segmentation. Character segmentation is a critical step of OCR system. Character segmentation is an operation that seeks to decompose an image of a sequence of characters

DOI: 10.15680/ijircce.2015.0306078

into sub images of individual symbols. It is depends on the script used in writing the document. A poor segmentation process produces misrecognition or rejection segmentation process carried after out only the pre processing of image [9]. There are various problem can be occur in character segmentation because all characters has not fixed size & shapes in handwritten document [8]. The problems in character segmentation can be divided into a variety of categories as following as:

i. Problem of broken characters
ii. Problem of overlapped characters
iii. Problem of Touching characters
iv. Problem of Skewed characters
v. Problem of irregular intensity with the character

IV. Feature Selection & Extraction

The features of the character are the distinct characteristics of that character which help to recognize it. For pattern recognition the features should be easily computed and robust. Two types of features used in pattern recognition. The first one has clear physical meaning, such as structural or statistical features. Another type of features has no physical meaning, but these features are treated as mapping features. The advantage of mapping features is that they can make classification easier. Statistical Features is a technique to machine intelligence which is based on statistical model in hand; one applies probability theory and decision theory to get an algorithm. Features are the measurements which represent the character such as size, shape and intensity. The statistical model one uses is crucially dependent on the choice of features. Hence it is useful to consider alternative representations of the same measurements (i.e. different features). For example, different representations of the character values in an image. Moment invariants have been frequently used as features for image processing [10], remote sensing, shape recognition and classification. Moments can its shape. Invariant shape recognition is provided characteristics of an object that uniquely represent performed by classification in the multidimensional moment invariant feature space. Several techniques have been developed that derive invariant features from moments for object recognition and representation. These techniques are distinguished by their moment definition, such as the type of data exploited and the method for deriving invariant values from the image moments. It was Hu that first set out the mathematical foundation for two-dimensional moment invariants and demonstrated their applications to shape recognition. These moment invariant values are invariant with respect to translation, scale and rotation of the shape.

Traditionally, moment invariants are computed based on the information provided by both the shape boundary and its interior region. The moments used to construct the moment invariants are defined in the continuous but for practical implementation they are computed in the discrete form. Given a function f(x,y), these regular moments are defined by:

$$\mathrm{M}_{pq} = \int\int x^{p} y^{q} f(x,y)dxdy \qquad \ldots(1)$$

Where, Mpq is the two-dimensional moment of the function f(x y) for p, q = 0,1, 2, …. The order of the moment is (p+q) where p and q are both natural numbers.

For implementation in digital form this becomes:

$$\mathrm{M}_{pq} = \sum_{\mathrm{X}}\sum_{\mathrm{Y}} x^{p} y^{q} f(x,y) \qquad \ldots(2)$$

V. Central Moment

A central moment is a moment of a probability distribution of random variable about the random variable's mean. The rth moment about any point is called a central moment; it is the expected value of a specified integer power of the deviation of the random variable from the mean. The various moments form one set of values by which the properties of a probability distribution can be usefully characterized. Central moments are used in preference to ordinary moments, computed in terms of deviations from the mean instead of from the zero, because the higher-order central moments relate only to the spread and shape of the distribution, rather than also to its location.

The central moment of a multivariate probability density function $P(x_1, x_2, \ldots)$can be similarly defined as

$$\mu_{m,n,\ldots} = \langle (x_1 - \langle x_1 \rangle)^m (x_2 - \langle x_2 \rangle)^n \cdots \rangle. \qquad ...(3)$$

Therefore,

$$\mu_{n,0,\ldots,0} = \mu_n. \qquad …(4)$$

For example,

$$\mu_{1,1} = -\mu'_{0,1}\,\mu'_{1,0} + \mu'_{1,1}$$

$$\mu_{2,1} = 2\,\mu'_{0,1}\,{\mu'_{1,0}}^2 - 2\,\mu'_{1,0}\,\mu'_{1,1} - \mu'_{0,1}\,\mu'_{2,0} + \mu'_{2,1}. \qquad …(5)$$

Similarly, the multivariate central moments can be expressed in terms of the multivariate cumulants.
For example,
µ1,1=K1,1
µ2,1= K 2,1
µ3,1=3 K 1,1 K 2,1 + K 3,1
µ4,1=6 K 2,0 K 2,1 + 4 K 1,1 K 3,0 + K 4,1
µ5,1=[15K22,0 + 10K2,1K3,0+10K2,0K3,1+5K1,1K4,0]+K5,1 …(6)
In this way 7th order Central moments features are extracted from the compound character. These values are combined with seven invariant moment features values.

VI. Classification and Recognition

The statistical technique has been most intensively studied and used in practice. More recently, neural network techniques and methods imported from statistical learning theory have been receiving increasing attention. The design of a recognition system requires careful attention to the following issues: definition of pattern classes, sensing environment, pattern representation, feature extraction and selection, cluster analysis, classifier design and learning, selection of training and test samples, and performance evaluation.

The invariant moments are well known to be invariant under translation, scaling, rotation and reflection. They are measures of the pixel distribution around the centre of gravity of the character and allow capturing the global character shape information. Traditionally, moment invariants are computed based on the information provided by both the shape boundary and its interior region. The moments used to construct the moment invariants are defined in the continuous but for practical implementation they are computed in the discrete form. Support vector machine was first introduced by V. Vapnik [8,13], it is simple and achieve good performance. The SVM is capable of learning a finite amount of data given for training. The principle of an SVM is to map the input data onto a higher dimensional feature space nonlinearly related to the input space and determine a separating hyper plane with maximum margin between the two classes in the feature space. [11].

VII. Result and Discussion

In this system we used combination of invariant moments and central moments. We used both types of compound character i.e with Shirorekha and without Shirorekha so that the proposed system should be implemented in any case. Hence the pre-processing task of detection and removal Shirorekha is reduced. This means the recognition time is minimized. The table 2 shows, average recognition rate of some of the sample compound characters.

Compound Character	Accuracy in %age
mya	93.33
nya	93.33
swa	90.00
tya	96.66
dhva	96.66
tva	90.00
stra	97.14
Avg	**93.87**

Table 2: Accuracy of Recognition of Some compound characters.

We have selected the samples with *'Shirorekha'* and without *'Shirorekha'* and also the combination of different characters. It is noted that the *'Shirorekha'* doesn't affect the recognition rate. As in case of 'tya' ,'dhav' the recognition rate is 96.66% for 'stra' it is 97.14% where as in case of 'tva' recognition rate is 90%. These compound characters are with *'Shirorekha'*. The proposed system with different kinds of compound characters, the average accuracy of recognition rate is 93.87%.

This recognition rate is increased by using some hybrid methods such as structural and statistical methods. Current research employs models not only of characters, but also words and phrases, and even entire documents, and powerful tools such as HMM[14], neural nets, contextual methods are being brought to bear. It is hoped that this comprehensive discussion will provide insight into the concepts involved, and perhaps provoke further advances in this area.

VIII. CONCLUSION AND FUTURE WORK

The simulation results showed that the proposed algorithm performs better with the total transmission energy metric than the maximum number of hops metric. The proposed algorithm provides energy efficient path for data transmission and maximizes the lifetime of entire network. As the performance of the proposed algorithm is analyzed between two metrics in future with some modifications in design considerations the performance of the proposed algorithm can be compared with other energy efficient algorithm. We have used very small network of 5 nodes, as number of nodes increases the complexity will increase. We can increase the number of nodes and analyze the performance.

REFERENCES

1. Veena Bansal and R. M. K. Sinha "Segmentation of touching characters in Devanagari", Indian Conference on Computer Vision, Graphics and Image Processing, Vol.-Issue.ICVGIP'98, pp. 371 - 376, 1998.
2. U. Pal, T. Wakabayashi and F. Kimura. "Comparative study of Devanagari Handwritten character recognition using different Features and Classifier", 978-0-7695-3725-2, IEEE, ICDAR, 2009.
3. Chavan S V , Kale K.V, Kazi M M and Rode Y S "Recognition of Handwritten Devanagari Compound Character A Moment Feature Based approach, International Journal of Machinne Intelligence, Vol. 5, Issue 1, pp 421-425, 2013.
4. V. S. Tapkir, S, D. Shelke "OCR For handwritten Marathi Script", International Journal of Scientific & Engineering Research, Vol. 3, Issue 8, 2012.
5. Amit Choudhary, Rahul Rishi, Savita Ahlawat "Off-Line Handwritten Character Recognition using Features Extracted from Binarization Technique.", Conference on Intelligent Systems and Control, AASRI Procedia Vol. 4, pp.306-312, 2013.
6. Priyanka Kamboj and Versha Rani "A Briff Study of Various Noise Model and Filtering Techniques", Journal of Global Research in Computer Science, Vol.4, Issue. 4, 2013.
7. Vneeta Rani, Dr.Vijay laxmi "Segmentation of Handwritten Text Document Written in Devanagari Script for Simple character, skewed character and broken character" International Journal of Computers & Technology, Vol. 8, Issue No 1, pp. 686-691, 2013.
8. Karbhari V. Kale, P D. Deshmukh, S V. Chavan, M M. Kazi and Yogesh S. Rode, "Zernike Moment Feature Extraction for Handwritten Devanagari (Marathi) Compound Character Recognition", International Journal of Advanced Research in Artificial Intelligence, (IJARAI), Vol. 3, Issue.1, 2014.

9. Vneeta Rani, Pankaj Kumar " Problems of character segmentation in Handwritten Text Documents written in Devanagari Script", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Vol..2, Issue 3, 2013.
10. P.E.Ajmire and S E Warkhede "Handwritten Marathi Character (Vowel) Recognition", Advances in Information Mining, ISSN: 0975–3265, Vol. 2, Issue 2, pp-11-13, 2010.
11. K V. Kale, P D. Deshmukh, S V. Chavan, M M. Kazi and Yogesh S. Rode, "Handwritten and Printed Devanagari Compound using Multiclass SVM Classifier with Orthogonal moment Feature", International Journal of Computer Applications (0975 – 8887) Volume 71– No.24, pp.31-37, 2013.
12. N. Joshi, G. Sita, A.G. Ramakrishnana, Deepu V, S.Mahdavnath, "Machine recognition of online handwritten Devanagari characters", Document Analysis and Recognition, Proceedings. Eighth International Conference, Vol.2, pp.1156 – 1160, 2005.
13. V. Vapnik., "The nature of statistical learning theory," Book published by Springer, N.Y. ISBN 0-387-94559-8, 1995.
14. U. Bhattacharya, S.K.Parui, B. Shaw, K. Bhattacharya "Neural Combination of ANN and HMM for handwritten Devanagari Numeral Recognition", Proceedings of the 10th IWFHR, pp. 613-618, 2006.

BIOGRAPHY

Prafulla Eknathrao Ajmire is a working as Associate Professor in the Department of Computer Science, G S Science, Arts & Commerce College, Khamgaon 444303 Dist. Buldana (M.S.), India. He received Master of Science.M.Sc.[Physics] from R.T.M. Nagpur University, Master of Science(M.S) in [Software System] degree in 1993 from BITS, Pilani, India. He received M.Phil [Comp.Sci.] degree in 2009 from S.G.B, Amravati University.

Dr. Rajiv V Dharaskar, Former Director, Disha Education Society (DIMAT - Disha Technical Campus), Raipur, C.G Former Director, MPGI Group of Institutions Integrated Campus, Nanded Former Professor and Head, PG Department of Computer Science and Engineering, G H Raisoni College of Engineering, a TEQIP II benefited Autonomous Institute, Nagpur.(India)

Dr. Vilas M Thakare, Professor and Head in Computer Science, Faculty of Engineering & Technology, Post Graduate Department of Computer Science, Sant Gadgebaba Amravati University, Amravati. Ph.D.(Computer Science), M.E. (Advance Electronics), M.Sc.(Applied Electronics), Diploma in Computer Management.