# Analysis of Feature Generated of Marathi Word

C.Namrata Mahender
Dept of CS and IT,
Dr.B.A.M.University,
Aurangabad

K.V.Kale
Dept of CS and IT,
Dr.B.A.M.University,
Aurangabad

## ABSTRACT

A continuous segmentation approach is used to get the isolated letters or pseudo character from the handwritten word for feature extraction. The preprocessing is done using morphological operation to remove noise, compress and provide smoother way for Feature extraction. To get better results Invariant moments (IM) are applied for feature extraction once preprocessing and segmentation is over because invariant moments are insensitive to translation, scale change, mirroring, and rotation. After segmentation it is often necessary to evaluate the quality of the image for that objective quality measure like mean and standard deviation are used and to see the relevance of our features for post-processing, correlation between the average of the group of similar pseudo-unit and sample of the similar pseudo-unit is calculated.

## Keywords

Feature extraction, segmentation, moments

## 1. INTRODUCTION

Writing which has been the most natural method of collecting, storing and transmitting information through centuries, now serves not only for the communication among humans, but also for the communication of humans and machines.

Since the advent of digital computer machine simulation of human reading has become the subject of intensive research. The research of handwritten recognition can be roughly divided into two categories global analysis and structured analysis. The Global or Holistic entails the recognition of the whole word by the use of identifying features as single indivisible entities and attempt to recognize them as whole bypassing the segmentation stage while words are suitable units for recognition. They are not a practical choice since each word has to be treated individually and data cannot be shared between word models, this implies a prohibitively large amount of training data.

Instead of using of whole word models, analytical approaches use sub word units such as characters or pseudo characters as the basic recognition units, requiring the segmentation of words into these units. Some of the problems and challenges, which are faced by researchers today, include: developing accurate segmentation, preprocessing feature extraction and classification techniques. For the first problem, segmentation the diverse styles and sizes of handwriting plays a large factor in the failure of current techniques .In some cases even a human being would not be able to segment handwriting containing continuous or mix characters, which is tightly packed together and illegible. These segmentation systems also have to deal with the variability due to variations found in handwriting amongst people but even difference can be seen in written documents of an individual too. This paper tries to address few problems relevant to segmentation preprocessing and quality of extracted features.

## 2. PREPROCESSING AND SEGMENTATION

Morphological operators are applied for preprocessing as it helps to reduce noise, compress size, and gets a processed image which is beneficial for further processing. After that an internal continuous segmentation is implied to get the appropriate pseudo character or units from the word image.

### 2.1 Preprocessing

In this system the processing of data is done via two different stages, so that we can have two sets of preprocessed data for further processing which will help to have comparative analysis of the results after feature extraction.

We are following Morphological based preprocessing, as it is a powerful tool for image enhancement, the Minkovsky morphological operations defined between the input image and a structuring element. Minkovsky operations replace convolution by logical operations. Two basic morphological operations are called dilation and erosion and are defined as follows: [2 3 6 13 14]

a). *Erosion :*
$I\oplus A=\{x : (Ax) \cap I\neq\Phi\}$,

b). *Dilation :*
$I\Theta A = \{x: (Ax) \subseteq I \}$,

The process of preprocessing done is as follows

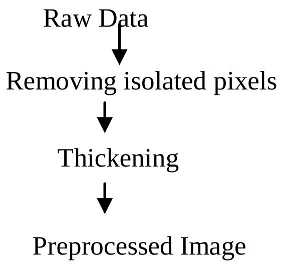

**Fig1: Preprocessed data**

### 2.2 Segmentation

Segmentation is an important stage, because the extent one can reach in separation of words, lines or characters directly affects the recognition rate of the script. There are two types of segmentation: external and internal. In this system internal segmentation is applied.

Internal Segmentation is the isolation of letters, especially, in cursively written words. Internal Segmentation is an operation that seeks to decompose an image of a sequence of characters into sub images of individual symbols. Explicit segmentation of cursive script into letters is still an unsolved problem. Methods for treating the problem have developed remarkably in the last decade and a variety of techniques have emerged. And in Devanagari script when half combined letters are there those are the point of problem while segmenting.[11 12 13 14]

The proposed segmented algorithm is a continuous segmentation of the given word with fixes counts. The data before used was resized to get the uniformity. The word image then segmented in two pass from left to right as well as from right to left.

Algorithm [Pass one]:- (From left to right segmentation. of the image)

Step 1:- Take the resized image of word.

Step 2:- Loop until c=xmax;

(xmax is the max value of x-axis of image,here it is 150.)

Step 3:- Initialize xmin=1,ymin=2;c=20.

Step 4:- Crop the image.

Step 5:- Resize the cropped image

(This is done to maintain the uniformity for extracting features).

Step 6:- Increment the counter by 5

Step7:- End loop.

In the pass two, the same procedure as in the pass one is applied just the direction is to the reverse of pass one i.e. from right to left.

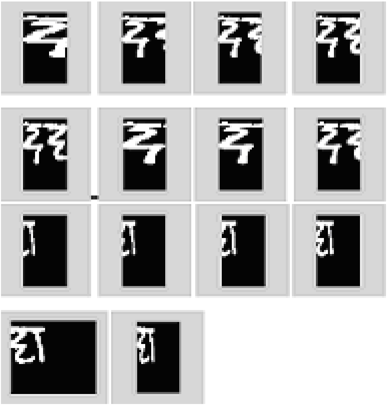

**Fig 2: Segmented using both the pass.**

## 3. FEATURE EXTRACTION

After the preprocessing is done on the two sets of data Invariant moments are applied on them for feature extraction.

### 3.1 Invariant Moments (IM)

The 2-D moments of order (p+q) of a digital image f(x,y) id defined as

$$m_{pq} = \sum_x \sum_y (x-\bar{x})^p (y-\bar{y})^q f(x,y)$$

for p,q = 0,1,2… where the summations are over the values of the spatial coordinates x and y spanning the image. The corresponding central moment is defined as [3]

$$\mu_{pq} = \sum_x \sum_y (x-\bar{x})^p (y-\bar{y})^q f(x,y)$$

Where

$$\bar{x} = \frac{m_{10}}{m_{00}} \quad \text{and} \quad \bar{y} = \frac{m_{01}}{m_{00}}$$

The normalized central moments of order (p+q) is defined as

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{pq}}$$

for p, q=0,1,2… where

$$\gamma = \frac{p+q}{2} + 1$$

where p+q=2,3,….

As set of seven 2-D moment invariants that are insensitive to translation, scale change, mirroring, and rotation can derive from these equations. They are as follows:

$$\phi_1 = \eta_{20} + \eta_{02}$$

$$\phi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2$$

$$\phi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2$$

$$\phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2$$

$$\phi_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2]$$
$$+ (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]$$

$$\phi_6 = (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] +$$
$$4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})$$

$$\phi_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2]$$
$$+ (3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]$$

IM are computationally simple but have some demerits like information redundancy, noise sensitivity.[5 6 13 ].

## 4. STATISTICAL ANALYSIS OF THE FEATURES

After segmentation, it is often necessary to evaluate the quality of the image. Therefore quality measures play an important role to evaluate the efficiency of segmentation technique in image processing application. For that an objective quality measure like mean and standard deviation are tested. [10]

Mean :

$$m = \sum_{i=0}^{L=1} z_i \, p(z_i)$$

Standard Deviation:

$$\sigma = \sqrt{\mu_2(z)}$$

**Table 1. Objective measures of few letters**

| ए | | क | |
|---|---|---|---|
| **Avg** | **Std** | **Avg** | **Std** |
| 0.9285 | 0.806 | 0.9374 | 0.0916 |
| 5.2317 | 1.3232 | 3.0038 | 0.7269 |
| 5.8049 | 0.5777 | 5.7087 | 1.4651 |
| 8.0487 | 1.4168 | 6.9262 | 0.9466 |
| 16.1901 | 2.4676 | 13.8750 | 2.1263 |
| 11.2793 | 1.9409 | 9.3519 | 1.6535 |
| 15.6227 | 2.1106 | 14.3811 | 2.0749 |

| दो | | न | | हा | |
|---|---|---|---|---|---|
| **Avg** | **Std** | **Avg** | **Std** | **Avg** | **std** |
| 0.8046 | 0.0784 | 0.6149 | 0.1053 | 1.1408 | 0.0757 |
| 3.5510 | 0.4323 | 2.5796 | 0.1053 | 5.3036 | 1.1457 |
| 4.0722 | 0.3493 | 4.7166 | 0.9932 | 5.3115 | 0.7274 |
| 6.5744 | 0.7651 | 7.1749 | 1.4218 | 9.0369 | 1.2714 |
| 12.6954 | 1.7943 | 12.6810 | 2.6557 | 16.8516 | 2.4411 |
| 8.4244 | 0.9724 | 8.6621 | 1.7385 | 12.7964 | 1.7954 |
| 13.0945 | 1.0526 | 12.9705 | 2.4112 | 17.2451 | 2.2013 |

Once the average and standard deviation is calculated we can test the pseudo-unit values with its corresponding average value calculated above if these mapped correlations are strong then we can take the average value of each pseudo unit as a separate class or vector which can be used for further processing like creating cluster or unique codebook for each such pseudo units. The table 2 shows the correlations between the average and the selected pseudo-unit of "KA" character in same way correlation for the remaining character is also tested.

**Table 2. Correlation of "ka" letter**

| | | **AVG** | **RESULT** |
|---|---|---|---|
| AVG | Pearson Correlation | 1 | .999** |
| | Sig.(2-tailed) | | .000 |
| | N | 7 | 7 |
| Result | Pearson Correlation | 999** | 1 |
| | Sig.(2-tailed) | .000 | |
| | N | 7 | 7 |

## 5. CONCLUSION

In this paper binary image are used and tried to help better feature extraction by preprocessing the image by using morphological operation. Morphological operation has given different angles to the raw image for further processing. A continuous segmentation is applied for getting the character or pseudo character from the word that provides a lot of combinations of pseudo-units for training and testing. Holistic technique is considered to get the overall feature of these segmented units in the form of invariant moments. Even to evaluate the quality of the segmented images objective quality measures are applied. The values we got from these features can be used for post processing as an input to the units based on statistical as well as neural networked based recognition system.

## 6. REFERENCES

[1] Verma, B., Blumenstein, M., Kukarni, S, 1998, "Recent Achievements in Off-line Handwriting Recognition Systems." International Conference on Computational Intelligence and Multimedia Applications, Australia.

[2] E.Kavallieratou N.Fakotakis G.Kokkinakis, 1999, "New Algorithms For Skewing Correction And Slant Removal On Word-Level (OCR)". The 6th IEEE International Conference on Electronics, Circuits and Systems, Vol 2, 1159 - 1162.

[3] Hasan Al-Rashaideh, 2006, "Preprocessing phase for Arabic Word Handwritten Recognition", Information Transmissions in Computer Networks, 11-19.

[4] S.Madhvanath and V.Govidaraju, 2001, "The Role of Holistic Paradigms in Handwritten Word Recognition", IEEE PAMI, Vol. 23, No.2, 149-164.

[5] Celebi, M.E. and Aslandogan Y.A., 2005, "A comparative study of three moment-based shape descriptors", Information Technology: Coding and Computing, International Conference, Vol 1, 788-793.

[6] Gonzalez, Woods, and Eddins, 2004, "Digital Image Processing Using Matlab", Prentice Hall, 470-472.

[7] K.Karacs and T.Roska, 2004, "Holistic Feature extraction from handwritten words on wave computers", Cellular Neural Networks and their Applications. Proceedings of the 8th IEEE international workshop Budapest, 364-369.

[8] S.Madhvanath, kim.G and Govindaraju,V, 1997," Chain code Processing for handwritten Word Recognition.IEEE

Transactions on Pattern Analysis and Machine Intelligence vol 21, 928- 932.

[9] Feng Pan and Mike Keane, 1994, "A New Set of Moment Invariants for Handwritten Numeral Recognition. Proceedings 1994, International Conference on Image Processing, Austin, Texas, USA, November 13-16, IEEE Computer Society.

[10] R.M. Bozinovic, and S.N. Srihari, 1980, "Off-Line Cursive Script Word Recognition", IEEE Trans.Pattern Analysis and Machine Intelligence, Vol. 11, 68-83.

[11] K.V. Kale,,R.R. Manza,S.S. Gornale,P.D. Deshmukh and Vikas Humbe, 2006, SWT Based Composite Method For Fingerprint Image Enhancement, 27-30.

[12] M. Blumenstein ,L.B. Verma,1999, Neural Based Solution For The Segmentation And Recognition Of Difficult Handwritten Words From Benchmark Database. Document Analysis and Recognition, 1999. ICDAR '99. proceedings of the Fifth International Conference,281-284

[13] C.Namrata Mahender and K.V. Kale, 2006, "Preprocessing of Offline Handwritten Marathi Word", Proceeding in National Conference on Image Processing, NCIP-2006, Pandharpur India, 24-25th December 2006.