A System for Off-line Oriya Handwritten Character Recognition using Curvature Feature

U. Pal[1], T. Wakabayashi[2] and F. Kimura[2]
[1]Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata-108, India
Email: umapada@isical.ac.in
[2]Graduate School of Engineering, Mie University, 1577 Kurimamachiya-cho, Tsu, Mie, Japan

Abstract

In this paper we present a system towards the recognition of off-line Oriya handwritten characters. Since most of the Oriya characters have curve-like stroke, we use curvature feature for the recognition purpose. To get the feature, at first, the input image is size normalized and segmented into 49×49 blocks. Curvature is then computed using bi-quadratic interpolation method and quantized into 3 levels according to concave, linear and convex regions. Next direction of gradient is quantized into 32 levels with π/16 intervals, and strength of the gradient is accumulated in each of the 32 directions and in each of the 3 curvature levels of every block. A spatial resolution is made to get 7×7 blocks from 49×49 blocks and a directional resolution is made to get 8 directions from 32 directions. Using curvature features for 3 levels we get 1176 (7×7 blocks × 8 directions × 3 levels) dimensional features. Finally using principal component analysis we reduce the dimension 1176 to 392 and this 392 dimensional feature vector is fed to a quadratic classifier for recognition. We tested 18190 samples of Oriya handwritten samples and obtained 94.60% accuracy from our proposed system.

1. Introduction

There are many pieces of work towards handwritten recognition of Roman, Japanese, Chinese and Arabic scripts, and various approaches have been proposed by the researchers towards handwritten character recognition [1]. Several pieces of research work are available on Indian printed characters but only a few attempts have been made towards the recognition of Indian off-line handwritten characters although there are many languages and scripts in India [2]. Some pieces of work have been done towards the recognition of Oriya printed characters [2-4] and handwritten Oriya numerals [5,6] but to the best of our knowledge there exists only one recent work on Oriya handwritten characters [7].

In this paper a quadratic classifier based scheme is proposed for unconstrained off-line Oriya handwritten character recognition. Since most of the Oriya characters are circular in nature, we employed curvature feature [8] with gradient feature for recognition and the dimension of the feature is 392. Detail of feature detection is discussed in Section 3. For the recognition we use a modified quadratic discriminant function and detail of this discriminant function is given in Section 4.

2. Oriya language and data collection

India is a multi-lingual and multi-script country and Oriya is one of the popular languages in India. Oriya is mainly used in the Indian state of Orissa. The Oriya script, by which Oriya language is written, is developed from the Kalinga script, one of the many descendents of the Brahmi script of ancient India.

The alphabet of the modern Oriya script consists of 11 vowels and 41 consonants. These characters are called *basic characters* and the basic characters of Oriya script are shown in Fig.1. Out of these 52 basic characters in Oriya, two characters are equal in shape (these two characters are marked by doted square box in Fig.1(b). For recognition, we consider these two characters as one class. Writing style of Oriya script is from left to right. Like other Indian scripts, concept of upper/lower case is absent in Oriya script. From Fig.1 it can be noted that most of the characters in Oriya is circular in nature and there is no horizontal line (like Shirorekha in Devnagari script) in the characters of this script.

In Oriya script a vowel following a consonant takes a modified shape. Depending on the vowel, its modified shape is placed at the left, right, both left and right, or bottom of the consonant. These modified shapes are called *modifiers or matra.* A consonant or a vowel 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. There are more than 200 compound characters in Oriya script [4] and in this paper we consider the recognition of off-line handwritten Oriya basic characters.

Main difficulty of any recognition system is shape similarity. In Oriya many characters have shape similarity. Examples of some similar shaped character groups are shown in Fig.2. From the figure it can be seen that shapes of two or more characters of a group is very similar and such shape similarity makes the recognition system more complex to get higher recognition rate.

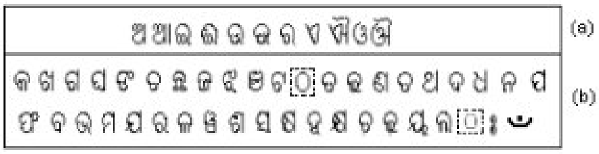


Fig.1. Basic characters of Oriya alphabet. (a) vowels and (b) consonants.

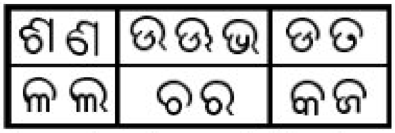


Fig.2. Examples of some similar shape Oriya characters.

Data collection for the present work has been done from different individuals of various professionals. We have considered 18190 data for the experiment of the proposed work. We used a flatbed scanner for digitization. Digitized images are in gray tone with 300 dpi and stored as TIF Format.

3. Feature extraction

Here we have used gradient and curvature features for our recognition purpose. For feature extraction at first the image is normalized and this normalized image is then segmented into 49 x 49 blocks. Compromising trade-off between accuracy and complexity, this block size is decided from the experiment. We apply Roberts filter on the image to obtain strength of gradient and direction of gradient image [7].

Curvature features can be computed in different ways and we computed curvature feature using bi-quadratic interpolation method because it gave better results according to the experiment of Shi et al. [8]. To get the features following steps are executed.

Step 1: The direction of gradient is quantized to 32 levels with $\pi/16$ intervals.

Step 2: The curvature c is computed by bi-quadratic interpolation method and quantized into 3 levels using a threshold t (for concave, linear and convex regions). For concave region $c \leq -t$, for linear region $-t < c < t$ and for convex region $c \geq t$. We assume t as 0.15 in our experiment.

Step 3: The strength of the gradient is accumulated in each of the 32 directions and in each of the 3 curvatures levels of each block to get 49x49 local joint spectra of directions and curvatures.

Step 4: A spatial and directional resolution is made as follows. A smoothing filter [1 4 6 4 1] is used to get 16 directions from 32 directions. On this resultant image, another smoothing filter [1 2 1] is used to get 8 directions from 16 directions. Further more, we use a 31 x 31 two-dimensional Gaussian-like filter to get smoothed 7 × 7 blocks from 49 x 49 blocks. So, we get 7×7×8 = 392 dimensional feature vector. Using curvature feature in 3 levels we get 392 × 3 =1176 dimensional features.

Step 5: Using principal component analysis we reduce 1176 dimensional feature vector to 392 dimensional feature vector and we fed this 392 dimensional feature vector to our classifier.

4. Character classifier

Recognition of characters in quadratic classifier [9] is carried out by using the following discriminant function:

$$g(X) = (N + N_0 + n - 1)\ln[1 + \frac{1}{N_0\sigma^2}[\|X - M\|^2 - \sum_{i=1}^{k}\frac{\lambda_i}{\lambda_i + \frac{N_0}{N}\sigma^2}\{\Phi_i^T(X - M)\}^2]] + \sum_{i=1}^{k}\ln(\lambda_i + \frac{N_0}{N}\sigma^2)$$

where X is the feature vector of an input character; M is a mean vector of samples; Φ_i^T is the ith eigen vector of the sample covariance matrix; λ_i is the ith eigen value of the sample covariance matrix; k is the number of eigen values considered here, n is the feature size; σ^2 is the initial estimation of a variance; N is the number of learning samples; and N_o is a confidence constant for σ and N_0 is considered as $2N/3$ from the experiment. We do not use all the eigen values and their respective eigen vectors for the classification. Here, we sort the eigen values in descending order and take first 180 (k=180) eigen values and their respective eigen vectors for classification.

5. Experimental results

Data used for the present work were collected from different individuals. We considered 18190 samples of Oriya basic characters (vowels as well as consonants) for the experiment of the proposed work. We have used 5-fold cross validation scheme for recognition result computation. Here database is divided into 5 subsets and testing is done on each subset using other four subsets in the learning. The recognition rates for all the test subsets are averaged to calculate recognition accuracy.

5.1 Global recognition results

From experiment we noted that the overall recognition accuracy of the proposed scheme was 94.60% when zero percent rejection was considered. 98.14%, 99.05%, 99.34% and 99.51% accuracy was obtained when we considered recognition result of two, three, four and five top choices, respectively, and no rejection was considered.

5.2 Rejection versus error rate computation

We also analyzed the rejection versus error rate of the classifier. We noted that 3.93%, 2.72%, 1.33% and 0.50% errors occur when the rejection rates are 2.99%, 6.00%, 12.02% and 19.77%, respectively. Rejection criteria of the proposed system is decided based on the 1st and 2nd values of the discriminant function *g(X).*

5.3 Confusing pair computation

We also noticed some confusing pairs of Oriya characters and their overall confusion rates are shown in Table 1. From the experiments we noticed that mainly similar shaped characters confused by the system at higher rate.

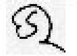
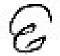
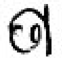
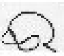
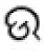
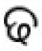
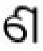
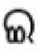
Table 1. Main confusing pairs of Oriya characters

Confusing character pairs		% of confusion (Overall)
		0.59%
		0.39%
		0.23%
		0.18%
		0.17%

5.4 Erroneous results

To get the idea about the samples where our system generates erroneous results, we provide some such samples in Table 2. Actual handwritten samples are shown in the first row of this table and the printed samples of their recognized class are shown in the respective columns of second row. Since the actual handwritten samples and recognized characters are very similar in shape we may think that these samples are recognized correctly. Unfortunately they are all miss-recognized. Actual class of each handwritten sample is shown in respective columns of the third row of the table.

Table 2. Erroneous samples

Actual Samples (Handwritten)				
Recognized as (Printed sample)				
Actual Class (Printed sample)				

5.5 Comparison of results

To the best of our knowledge there exists only one recent work on off-line handwritten Oriya characters and that is done by us [7]. We obtained 91.11% accuracy from that experiment and it was tested on 9556 Oriya handwritten character samples. From this current method we obtained 94.60% accuracy from 18190 Oriya handwritten samples. Please note that 9556 data samples used in [7] are included in this present database of size 18190.

References

[1] R. Plamondon and S. N. Srihari, "On-line and off-line handwritten recognition: A comprehensive survey", IEEE Trans. on PAMI, Vol.22, pp. 62-84, 2000.

[2] U. Pal and B. B. Chaudhuri, "Indian Script Character Recognition: A Survey", Pattern Recognition, Vol.37, pp.1887-1899, 2004.

[3] S. Mohanty, "Pattern recognition in alphabets of Oriya language using Kohonen neural network" IJPRAI, Vol. 12m pp. 1007-1015, 1998.

[4] B. B. Chaudhuri, U. Pal and M. Mitra, "Automatic recognition of printed Oriya script", Sadhana, Vol.27, part 1. pp.23-34, February 2002.

[5] K. Roy, T. Pal, U. Pal and F. Kimura, "Oriya handwritten numeral recognition system" In Proc. 8th ICDAR, pp. 770-774, 2005.

[6] T. K. Bhowmick, S. K. Parui, U. Bhattacharya, B. Shaw, "An HMM based recognition scheme for handwritten Oriya numerals", In Proc. 9th ICIT, pp. 105-110, 2006.

[7] U. Pal, N. Sharma, and F. Kimura, "Oriya offline handwritten character recognition", In Proc. International Conference on Advances in Pattern Recognition, pp.123-128, 2007.

[8] M. Shi, Y. Fujisawa, T.Wakabayashi, and F. Kimura, "Handwritten numeral recognition using gradient and curvature of gray scale images", Pattern Recognition, Vol.35, pp.2051-2059, 2000.

[9] F Kimura, K, Takashina, S. Tsuruoka and Y. Miyake, "Modified quadratic discriminant function and the application to Chinese character recognition", IEEE Trans. on PAMI, Vol. 9, pp 149-153, 1987.