OHCS: A Database for Handwritten Atomic Odia Character Recognition

Ramesh Kumar Mohapatra, Tusar Kanti Mishra, Sandeep Panda, Banshidhar Majhi
Computer Science and Engineering
National Institute of Technology
Rourkela, India-769008
Email: mohapatrark@nitrkl.ac.in, tusar.k.mishra@gmail.com, sandeep.panda91@gmail.com, bmajhi@nitrkl.ac.in

***Abstract*—In this paper, a complete database of handwritten atomic Odia characters is suggested. The first version of the database has been modeled and named OHCSv1.0 (Odia handwritten character set). The database comprises of 17,100 transcribed characters, each collected twice from 150 unique people at different point of time. Each character has 300 number of occurrences. The character images are standardized to a size of 64×64 pixels. A novel framework for perceiving transcribed Odia characters from this database has also been proposed. The character images are gathered into various groups in view of their shape components utilizing an incremental spectral clustering algorithm. During testing, affinity of probe character to a cluster is first decided. Subsequently, the trained classifier recognizes the character inside the cluster. Suitable simulation has been carried out to validate the scheme.**

I. Introduction

Optical Character Recognition (OCR) has been a prevalent examination territory for a long time on account of its different application possibilities, for example, bank check processing, postal automation, documents analysis, archival of historical documents and so forth. A few logical looks into have been completed to perceive characters in distinctive dialects like *English, Chinese, Arabic, Urdu, Devnagari, Tamil, Telgu, Gujurati, Bengali, Assamese, Odia* [1], [3], [6]–[9]. There exist some standard databases for English numeral and character, Devnagari character, Chinese character, and few others. Therefore, to institutionalize and analyze exploration results, numerous databases in the manually written acknowledgement space have been accumulated and utilized as a part of different dialects and applications. However, only a few studies are attested on handwritten characters of the Odia scripts. Recently, some efforts have been reported in literature for Odia characters recognition. These recognition schemes are based on the curvature feature [2], the F-ratio based weighted feature extraction for similar shape character recognition [4], Odia character recognition using Kohonen neural network [10], and multi font Odia character recognition using curvelet transform [11]. However, these schemes are simulated by researchers through their own generated datasets with few samples as there exists no such standard dataset on Odia handwritten character set publicly in a sufficient volume. Hence, results standardization and comparative analysis have become a difficult task. As a part of the research on recognition of Odia characters, we developed a database and named it as Odia Handwritten Character Samples database version 1.0 (OHCS v1.0). The database consists of 17,100 handwritten characters collected from 150 different individuals at different points of time. The characters have been size-normalized and centred in a fixed-size of 64×64. This database has been developed at the Computer Vision and Pattern Recognition Research Laboratory of the National Institute of Technology, Rourkela, India.

A recognition scheme based on clustering technique has been proposed for character recognition for this database. Incremental spectral clustering [12] has been used to group the alphabets into cluster of similar shape. Three different recognition schemes are used for comparative study and select the best classifier for the proposed method. The classifier that provides the highest accuracy is selected for comparison with standard schemes and compare the efficacy of the proposed scheme.

II. An Overview of the Odia Script

Odia is the vital and local dialect of the Indian states of Odisha (Former Orissa). It is a dominating language, where local speakers involve 80% of the populace and rest 20% originates from different parts of West Bengal, Jharkhand, Chhattisgarh, and Andhra Pradesh. It is the official language of Odisha state and is the second dialect of Jharkhand. Odia is the sixth Indian language to be recorded as a classical language in India, on the premise of having a long abstract history and not having acquired broad topic from different dialects. The Odia script is created from the Kalinga script, one of the descendants of Brahmi script of antiquated India.Odia writing system is unique in relation to some other territorial dialects in India. The present day Odia script comprises of 11 vowels, 36 consonants, and 10 numerals.

The structure of Odia characters are mostly round-formed like the Devanagari and Bengali characters, however the later dialects have a flat line on the top (called *Sirorekha*) which is truant in Odia. Odia language is not case touchy. The end of a sentence is stamped by a vertical line ('|') instead of a period ('.'). One of the significant attributes of Odia rudimentary characters is that most of their upper 33% is round shaped and a subset of them have a vertical straight line at their furthest right part. In free-form transcribed character identification, the characters are thought to be composed intelligibly permitting littler variety fit as a fiddle of a character.

This is a pre-print version of a paper accepted to NCVPRIPG 2015, IIT Patna

Fig. 1: Handwritten Odia isolated character set.

III. OHCS v1.0

The proposed model comprises of data acquisition and essential strides of preprocessing steps like image binarization, skew angle detection and correction, text line segmentation followed by character segmentation. At long last, each secluded character is standardized and put away into the repository. Our database of isolated Odia handwritten character samples are collected from 150 distinct people. These writers are selected from various age, gender, and educational background groups. The samples are collected by asking the informants to write on a formatted sheet containing the Odia character set. Along these lines, a person gives an aggregate of 57 test images. We chose an input page layout that makes the segmentation simple, to stay away from the perplexing issue of document segmentation in characters. We composed a structure containing the printed dis-engaged Odia characters and empty space just below for each isolated printed character (as references for the users) which has to be filled by a writer. A digital note-maker device (i-Ball TakeNote) has been utilized for the purpose to avoid noise due dust and spilling of fluids. The pictures are put away in *.bmp* format to retain maximum image information. A sample handwritten Odia character set is shown in Figure 1.

Binarization is carried out for each digital page for reducing the space complexity. Skew correction is applied projecting the page at several angles, and deciding the change in the number of black pixels per projected line. After getting the skew angle, a page slope correction was performed on the digital page. Original isolated characters from the digital page are extracted by utilizing bounding box method. After the digitizing the collected forms, we process and segment them into isolated characters using an automated system built for the purpose. In reality, a page incline adjustment is performed naturally utilizing the Hough transform to appraise the skew edge and correct the skewness of the examined images. Next, we built up a propelled horizontal projection technique for automatic extraction of original isolated characters from the digital page.

For storing all standardized images we have created folders for each class of data and we assign a label to each class. All images of one character are assigned one label. So, we obtained 57 such classes and in each class there are 300 normalized character images. These folders are named by the numeral values from 1 to 57 for each class of characters according to the Table I.

TABLE I: List of files with size in bytes.

File name	Size in bytes
OHCS-train-samples.gz	10,246
OHCS-train-labels.gz	4,980
OHCS-test-samples.gz	3,614
OHCS-test-labels.gz	1,016

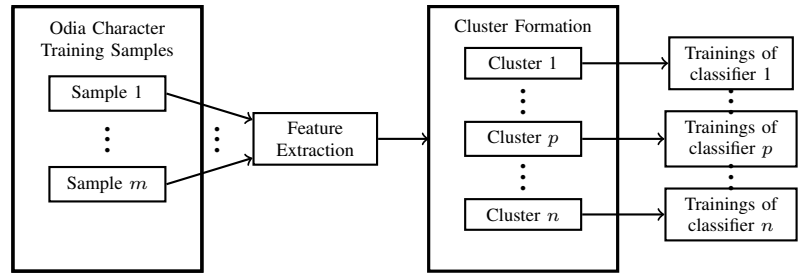
The OHCS 1.0 database of transcribed Odia basic isolated characters has a training set of 8,550 illustrations, and a test set of 8,550 cases. The characters have been size-normalized and centred in a fixed-size of 64×64. It is a decent database for individuals who need to have a go at learning techniques and pattern recognition methods on real-world data while spending negligible endeavors on preprocessing and organizing the data.

A. Proposed Recognition Framework

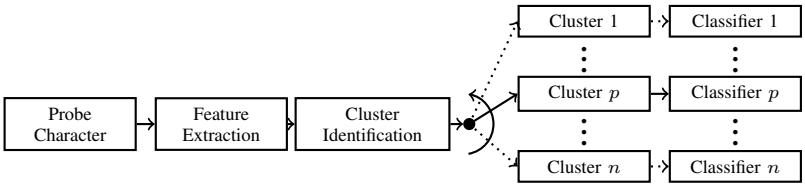
A framework to recognize Odia characters has been proposed and evaluated on the OHCS 1.0 database. In this section, we propose an efficient framework to recognize the Odia scripts.

Figure 2 shows the proposed recognition framework. The preprocessed character image is subjected to feature extraction, where four different histogram primitives are generated. The first histogram counts frequency of foreground pixel (black) on each vertical scan line and similarly the second histogram counts the frequency of foreground pixel on each horizontal scan line. We have taken 64 number of scan lines on each histogram for generating two vectors each of size 1×64. The third histogram is constructed in a horizontal manner, which gives the number of background pixels encountered prior to the occurrence of the first foreground pixel along the same line. The fourth histogram, is constructed using vertical scan lines and it gives the background pixels count in the vertical direction. The later two histogram also generate two vectors each of size 1×64. All the vectors are concatenated to generate a final feature vector of size 1×256. The first two feature vectors contain the shape information of each character from bottom to top and from left to right. The later two histograms keep information of the shape of the character and its curvatures along the horizontal and vertical direction where the first two histograms are made. Figure 3 shows all the instances of horizontal and vertical histograms for foreground and background pixel frequencies of the first alphabet of the Odia script ଅ('ah'). Figure 3 shows the horizontal and vertical histograms of the same character.

In the training phase, feature vectors are generated from the entire Odia character set, which are subsequently grouped into n clusters depending on their histogram features that corresponds to the shape information only. Three unique classifiers, to be specific, back propagation neural network (BPNN), k-nearest neighbor approach (KNN), and support vector machine (SVM) are trained on the cluster bins. While



(a) Training and clustering.



(b) Testing.

Fig. 2: The proposed Odia character recognition framework.

testing for a character image, it is passed through two phases for detecting its class label. Initially it is tested for the bin classification. At the second level, it is fed to all the three classifiers separately for generating the target class. We should be aware that, the number of characters in each cluster is much less than 57, assuming there are no empty bins. Thus the number of output classes inside a particular cluster, is always less than 57 which helps in improving both accuracy and computational time.

IV. Simulation Results and Analysis

Simulations are carried out to validate the proposed recognition scheme. The simulation consists of two parts as described below.

A. Experiment 1 : Clustering

Distance of the probe character image to each cluster center is computed. Based on the minimum distance, it is classified into the corresponding bin. An instance of these distances for an Odia test character ଅ('ah') is shown in Figure 4. As can be seen, then character has been successfully classified into its correct cluster. The number of probes successfully classified into their corresponding bins are recorded and the overall penetration rate is calculated. The penetration rate for all the clusters put together is obtained as 83.75%.

B. Experiment 2 : Recognition

Finer level classification of the semi-classified probe characters is carried out in this phase. The classifiers are trained to recognize the characters in the clusters. To train the classifiers, we have used the same global histogram features that have been discussed in section III-A. A total of 7520 samples have been used for the purpose of training the three classifiers. Each image has a feature vector of size of 1×256 which is the feature vector obtained from its histograms variants. Comparison among the three classifiers (BPNN, KNN, SVM) is drawn among them to estimate the best selection for the task of recognition. Results are presented in Table II.

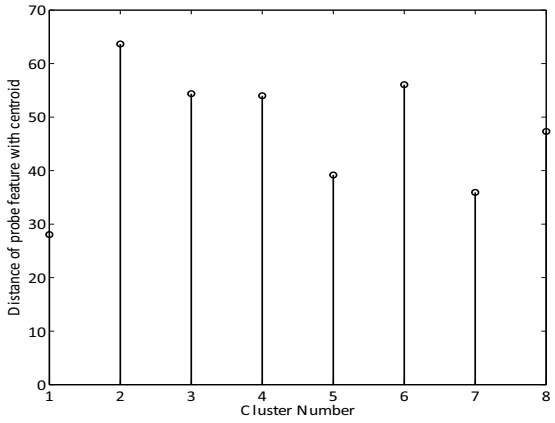


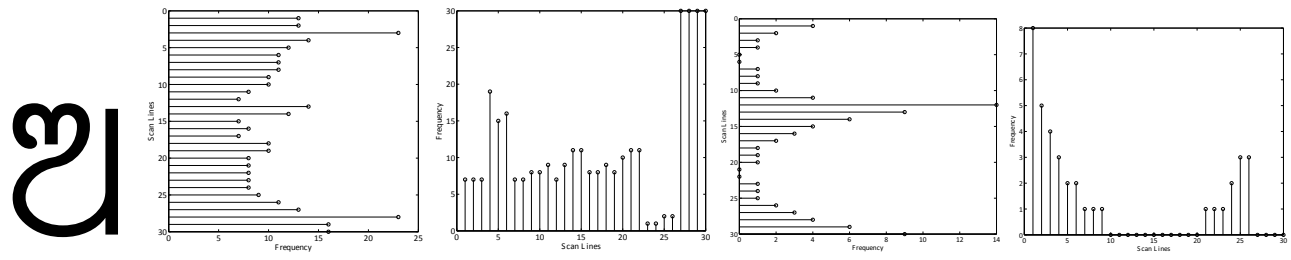
Fig. 4: Distance of the first alphabet from each cluster.

TABLE II: Performance comparison of various classifiers with varying grid size.

Classifier	Accuracy	TOC (Time of Computation)
BPNN	79.25	0.12
KNN	81	0.11
SVM	83.75	0.13

As seen from this table, the SVM classifier provides a higher rate of accuracy as compared to the other two classifiers. SVM can model highly non-linear relationships due to its higher dimensional plotting of feature points and use of maximum separating hyperplane. It can also be observed that the SVM takes slightly more time for computation as compared to the others. However, it is still a better choice than the others so far as the rate of recognition is concerned.

The proposed recognition frame work is compared with other competent recognition schemes. These schemes have



(a) (b) (c) (d) (e)

Fig. 3: Feature extraction. (a) First alphabet of the Odia script, (b) Horizontal histogram of Foreground Pixel Frequency, (c) Vertical histogram of Foreground Pixel Frequency, (d) Horizontal histogram of Background Pixel Count and (e) Vertical histogram of Background Pixel Count.

been simulated with the features suggested by respective researchers. Training is carried with 8550 samples and testing is carried out with the other 8550 samples which are randomized for 10 separate instances of the experiments to provide an unbiased output. A comparative study of the rates of accuracy are presented in Table III. The results show that the proposed scheme outperforms the state of the art schemes in terms of recognition rates. A k-fold cross-validation (k=5) is carried out on the test samples and rates of accuracies are plotted with an increment of 10 samples per class increment to the test set. Results thus obtained are shown in Figure 5.

V. Conclusion

In this paper, first of its kind database for handwritten Odia atomic characters has been presented. We propose a robust scheme for handwritten character recognition of the Odia language. The scheme utilizes vertical and horizontal histograms of the characters with both global and local approaches, to extract features. Cluster sets are constructed in coarser level using the histogram features to reduce computational overhead during classification. In finer level of classification, three different classifiers are evaluated. It is found that, the SVM classifier shows maximum rate of accuracy as compared to KNN and BPNN. The suggested recognition scheme outperforms other state-of-the-art schemes. As the future work, initiatives are now taken to create a database of handwritten compound Odia characters.

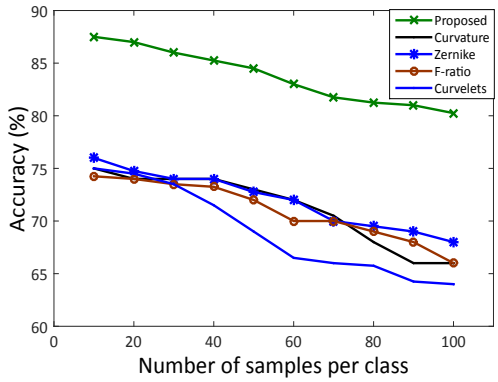


Fig. 5: Overall comparison of the proposed scheme with other competent scheme.

TABLE III: Computational overheads for different schemes.

Scheme	Accuracy	TOC	TOC without Clustering
F-ratio [4]	71.25	0.11	0.45
Curvelets [11]	71	0.12	0.47
Curvature [2]	69	0.12	0.46
Zernike Moments [5]	72	0.13	0.46
Proposed Scheme	83.75	0.13	0.47

References

[1] U. Pal, R. Jayadevan, N. Sharma, Handwriting Recognition in Indian Regional Scripts: A Survey of Offline Techniques, ACM Transactions on Asian Language Information Processing, Vol. 11, No. 1, pp. 1 - 35, 2012.

[2] U. Pal, T. Wakabayashi, F. Kimura, A System for Off-Line Odia Handwritten Character Recognition Using Curvature Feature, 10th International Conference on Information Technology, pp. 227 - 229, 2007.

[3] B.B. Chaudhuri, U. Pal, M. Mitra, Automatic recognition of printed Odia script, Sixth International Conference on Document Analysis and Recognition, pp. 795 - 799, 2001.

[4] T. Wakabayashi, U. Pal, F. Kimura, and Y. Miyake, F-ratio based weighted feature extraction for similar shape character recognition,International Conference on Document Analysis and Research, pp. 196 200, 2009.

[5] A. Khotanzad, Y.H. Hong, Invariant image recognition by zernike moments, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.12, pp. 489 497, 1990.

[6] G. Siromoney, R. Chandrasekaran, M. Chandrasekaran, Computer recognition of printed Tamil characters, Pattern Recognition, Vol. 10, Number. 4, pp. 243 247, 1978.

[7] G. S. Lehal and C. Singh, and M. Chandrasekaran, A Gurmukhi Script Recognition System, International Conference on Pattern Recognition, Vol. 4, pp. II-557 II-560, 2000.

[8] S. Bag, G. Harit and P. Bhowmick, Recognition of Bangla compound characters using structural decomposition, Pattern Recognition, Vol. 47, Number. 3, pp. 1187 1201, 2014.

[9] A. A. Desai, Gujarati handwritten numeral optical character reorganization through neural network, Pattern Recognition, Vol. 43, Number. 7, pp. 2582 2589, 2010.

[10] S. Mohanty, Pattern Recognition in Alphabets of Oriya Language using Kohonen Neural Network, International Journal of Pattern Recognition and Artificial Intelligence, Vol. 12, Number. 7, pp. 1007 1114, 1998.

[11] S. Nigam, A. Khare, Multifont Oriya Character Recognition Using Curvelet Transform, Information Systems for Indian Languages, Communications in Computer and Information Science Volume 139, pp 150-156, 2011.

[12] U. V. Luxburg, A tutorial on spectral clustering, Statistics and Computing, Vol. 17, Number. 4, pp. 395 416, 2007.