IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.7, Special Issue.1, Jan 2019 E-ISSN: 2347-2693

A Hybrid Recognition System of Handwritten OlChiki Character and Digit

SumantaDaw[1*], Abhoy Chand Mondal[2]

[1]Department of CSE, Hooghly Engineering & Technology College, Hooghly
[2]Department of CSE, The University of Burdwan, Burdwan

[*]Corresponding Author: sumanta.daw@hetc.ac.in

Available online at: www.ijcseonline.org

Abstract- The process of recognizing scanned documents or machine printed documents using automated tools are used in different real life domains. Designing a method with cent percent accuracy of character recognition is a challenging and unachievable task. Presence of noise, distinct styles of font under real time environment makes character recognition more difficult.

In this paper we describe recognition of handwritten basic characters of OlChiki script, used by more than 10 million tribal people in India mostly from Assam, Bengal, Bihar, Odisha and Jharkhand. There are 30 basic characters and 10 numeral digits in OlChiki and we have used a dataset of 10000 handwritten isolated character samples written by 50 persons. Samples in this dataset are composed of one stroke. Curvelet and Geometry based feature extraction has been used for comparison of performance. Strokes are recognized dynamically by using KNN and SVM classifier together. We have received an encouraging recognition result of 87% accuracy.

Keywords- OlChiki Script, Basic Character Recognition, Curvelet based Feature, Geometry based Feature, KNN Classifier, SVM Classifier.

I. INTRODUCTION

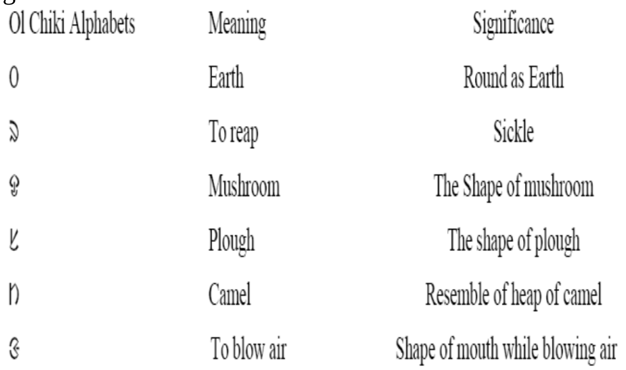
The OlChiki script, also called OlCemet', OlChiki, or simply Ol, was invented by Pandit Raghunath Murmu in the first half of the 20th century, spoken by ten million Santali people, mostly in India with a few in Nepal and Bangladesh.

The Santali is the language spoken by one of the largest indigenous socio-linguistic group in the Indian subcontinent and its speakers presently resides mainly in the Indian states of Assam, Bihar, Jharkhand, Orissa and West Bengal, and also in Bangladesh and Nepal. Santali language belongs to the Munda group of Austro-Asiatric family of languages. The Santali language does carry Santals' heritage and traditions in the form of folklore and tales since time immemorial. Currently Santali language is written in five different scripts, viz. OlChiki, Bengali, Devnagari, Oriya and Roman Scripts. Out of these scripts, it is only the OlChiki script that has been devised purely for writing Santali, in particular, and Munda group of languages, in general.

After the invention of OlChiki script during 1930s, a large number of books have been written by various authors in Santali using OlChiki script. Therefore, it is of great importance to preserve Santali language in its authentic form, and evidently the script plays an important role to perpetuate a language in its true form among the members of its speakers. In other words, the propagating OlChiki for the development of Santali language is the only way of carrying forward its heritage. In the year of December, 2002 OlChiki script has been computerized with ASCII Code.

Novelty of OlChiki Script:

One of the interesting features of OlChiki script is that it makes use of signs and symbols long familiar to the Santals. Letters of OlChiki script are also derived from the physical environment, likes – hills, rivers, trees, birds, bees, plough, sickle etc. [1, 2]. A partial list is given below in Figure 1 to show some OlChiki Alphabets with their meaning and significance.

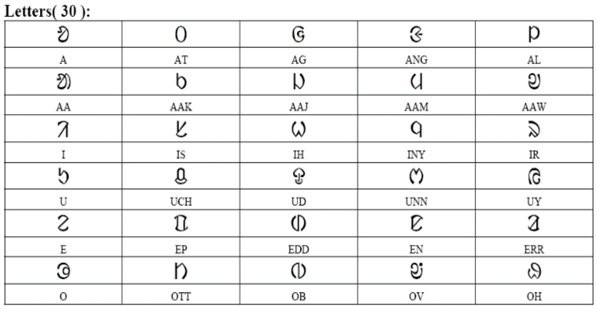


Ol Chiki Alphabets	Meaning	Significance
ᱚ	Earth	Round as Earth
ᱨ	To reap	Sickle
ᱩ	Mushroom	The Shape of mushroom
ᱥ	Plough	The shape of plough
ᱦ	Camel	Resemble of heap of camel
ᱝ	To blow air	Shape of mouth while blowing air

Figure 1: OlChiki alphabets with their meaning and significance

Usefulness of Recognition System of OlChiki:

Santals are not able to harness the power of the great global phenomenon known as the 'Computer'. This paper focuses on the recognition of OlChiki characters and as well as of OlChiki script for the Document Processing in near future. In OlChiki script, there are 30 characters (Figure 2a (Printed), 2b (Handwritten)) and 10 digits (Fig. 3a (Printed), 3b (Handwritten)). OlChiki Printed [7] and Hand written characters are not similar in shape. This paper concentrates on the recognition of the basic handwritten characters of OlChiki.



Letters(30):

ᱚ	ᱛ	ᱜ	ᱝ	ᱞ
A	AT	AG	ANG	AL
ᱟ	ᱠ	ᱡ	ᱢ	ᱣ
AA	AAK	AAJ	AAM	AAW
ᱤ	ᱥ	ᱦ	ᱧ	ᱨ
I	IS	IH	INY	IR
ᱩ	ᱪ	ᱫ	ᱬ	ᱭ
U	UCH	UD	UNN	UY
ᱮ	ᱯ	ᱰ	ᱱ	ᱲ
E	EP	EDD	EN	ERR
ᱳ	ᱴ	ᱵ	ᱶ	ᱷ
O	OTT	OB	OV	OH

Figure 2a: OlChiki basic Printed Characters

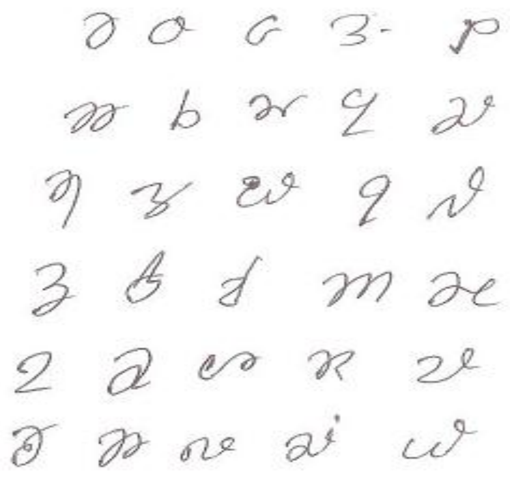
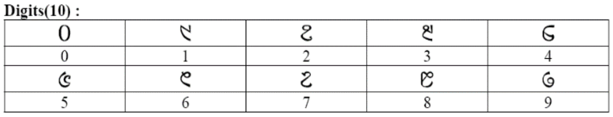


Figure 2b: OlChiki Basic Handwritten Characters



Digits(10) :

᱐	᱑	᱒	᱓	᱔
0	1	2	3	4
᱕	᱖	᱗	᱘	᱙
5	6	7	8	9

Figure3a: OlChiki Digits (Printed)

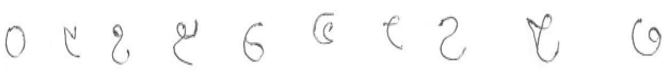


Figure 3b: OlChiki Digits (Handwritten)

Steps involved in Character Recognition System:

In this paper to recognize the Basic OlChiki handwritten characters and digits the following important steps has been used i.e. Feature Extraction, Classification and Handwritten Character Recognition. The feature extraction technique based on the digital Curvelet and geometry based transformation is used. Recognition accuracy is increased by using hybrid classifier with KNN and SVM. Finally, the output values of these classifiers are used to make the final decision.

II. HANDWRITTEN CHARACTER RECOGNITION

As with most pattern recognition tasks, OlChiki character recognition also involves the following important steps: Pre-processing, feature extraction and classification.

Pre-processing:

In handwritten OCR, document image is first captured by a scanner. Pre-processing of the scanned image significantly improves the efficiency of the document analysis process. Binarization is an important pre-processing step which converts gray image into a binary image. However, the existing global binarization methods are generally not suitable for a non-uniformly illuminated document as the threshold value is same for the whole document image but not the illumination. Therefore, locally adaptive binarization techniques are usually better suited in case of non-uniformly illuminated and degraded documents. In our work, all scanned images are assumed to be uniformly illuminated. Accordingly, the global thresholding method proposed by Otsu [3] is used without any problem.

Feature Extraction

After pre-processing the document image, features are extracted. These feature values are subsequently fed to the classifier.

- *Curvelet transform based feature extraction:*

Some earlier proposed character features [6] are used for recognition, i.e. directed chain code, intersection, shadow feature, chain code histogram and straight line fitting features, gradient and curvature information. In handwritten text, one common but important feature is the orientation of the text written by the writer. Also, for large set of characters, as in Bangla, Devnagari, etc., automatic curve matching is highly useful. Accordingly, Curvelet transform for extracting features from handwritten OlChiki character has been used.

Curvelet represents edges and singularities along curves more precisely with the needle shaped basis elements. The elements possess super directional sensitivity and capability to capture smooth contours. Since curvelet are two dimensional waveforms that provide a new architecture for multi-scale analysis, they may be used to distinguish similar appearing characters better. In our proposed curvelet-based feature extraction, the characters in a document image are first extracted using conventional methods. Each character sample is then cropped and resized so as to fit within a frame of standard width and height. Following this, the digital Curvelet transform at a single scale is applied to each of the character samples in the document to obtain Curvelet feature coefficients characterizing. In this work, we compute 1024 (i.e., 32 × 32) feature coefficients.

Algorithm 1: Curvelet Transform
1. Sub-band decomposition: The image is divided into resolution layers where each layer contains details of different frequencies.
2. Smooth Partitioning: Each sub-band is smoothly windowed into “squares” of an appropriate scale.
3. Renormalization: Each resulting square is renormalized to unit square.
4. Ridgelet analysis: Each square is analysed in the ortho-ridgelet system.

- *Character Geometry based feature extraction:*

In this feature extraction technique, the geometric features of the character contour are extracted. These features are based on the basic line types that form the character skeleton.

Algorithm 2: Character Geometry Feature Extraction
1. Universe of Discourse: First, universe of discourse is selected because the features extracted from the character image include the positions of different line segments in the character image.

2. Zoning: The image is divided into windows of equal size and feature extraction is applied to each individual zone rather than the whole image. In our work, the image was partitioned into 9 equal sized windows.

3. Starters, Intersections and Minor Starters: To extract different line segments in a particular zone, the entire skeleton in that zone should be traversed. For this purpose, certain pixels in the character skeleton are defined as starters, intersections and minor starters.

4. Character traversal: Character traversal starts after zoning by which line segments in each zone are extracted. First, the starters and intersections in a zone are identified and then populated in a list. Algorithm starts by considering the starter list. Once all the starters are processed, minor starters obtained along the course of traversal are processed. The positions of pixels in each of the line segments obtained during this process are stored. Once all the pixels in the image are visited, the algorithm stops.

5. Distinguishing line segments: After all the line segments in the image are extracted, they are classified into any one of the following line-types – Horizontal line, Vertical line, Right-diagonal line, or Left-diagonal line.

6. Feature Extraction: After the line type of each segment is determined, feature vector is formed based on this information which include the number and the normalized length of the four different types of lines in each zone. The normalized length of a line is given as Normalized Length =No. of Line Pixels/No. of Zone Pixels.

After zonal feature extraction, certain features are extracted for the entire image based on the regional properties, viz., Euler number, regional area, and eccentricity.

III.CLASSIFICATION

The main task of classification is to use the feature vectors provided by the feature extraction algorithms to assign the object to a category. In our work, we used SVM and k-NN hybrid classifier for OlChiki characters recognition.

Support Vector Machine
Support vector machine (SVM) was developed by Vapnik [4] and is an extensively used tool for pattern recognition due to its many attractive features and promising empirical performance, especially in classification and nonlinear function estimation. SVMs are used for time series prediction and are comparable to radial basis function network. The performance of SVM classification is based on the choice of kernel function and the penalty parameter C. In our work, we used SVM classifier with Radial Based Kernel for the classification of isolated characters. The RBF kernel maps nonlinear samples into a higher dimensional space, and thus can handle the case when the relation between class labels and attributes is nonlinear. SVM efficiently works for spaces with high dimension. SVM is versatile and flexible to use.

K-Nearest Neighbour Classifier
The k-Nearest Neighbour (k-NN) classifier classifies an unknown sample based on the known classification of its neighbours [5]. Given an unknown data, the k-nearest neighbour classifier searches the pattern space for the k training data that are closest to the unknown data. These k training tuples are the k “nearest neighbours’” of the unknown data. “Closeness” is defined in terms of a distance metric, such as the Euclidean distance. Typically, we normalize the values of each attribute. This helps to prevent attributes with initially large ranges from out weighting attributes with initially smaller ranges (such as binary attributes).Min-max normalization, for example, may be used to transform a value of a numeric attribute. The advantage of using KNN classifier is that it is very easy and working good for simple problems of recognition.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

Since standard benchmark data-set for handwritten characters are not available for OlChiki, we collected a data-set of

10000 samples of 30 OlChiki characters and 10 digits written by different writers to evaluate and compare the performance of different feature extracted with different classifiers. Here we have used K-Fold cross validation technique for achieving a good result. That K-Fold cross validation is a procedure used to estimate the skill of the model on new data. In this method, we split the data-set into k number of subsets (known as folds) then we perform training on the all the subsets but leave one (k-1) subset for the evaluation of the trained model. In this method, we iterate k times with a different subset reserved for testing purpose each time. Feature extraction was done on each sample using both Curvelet Transform and Character Geometry at a single scale. The Curvelet-based feature vectors obtained had a dimensionality of 1024. Principal component analysis of the coefficients was performed to reduce the dimension of the feature vectors to 190. In Character Geometry-based algorithm, we extracted 86-dimensional feature vectors and so there was no need to reduce their dimension. KNN classifier with Curvelet based features is providing a promising result than SVM classifier. But some OlChiki characters are not recognizing by KNN although it is a simple classifier where most of the OlChiki characters are single stroke based. For that reason we have generated two distinguish sets of OlChiki characters based on strokes. First set is recognizing with KNN classifier and other one by SVM (Figure 4 shows that). We evaluated and compared both the types of features in two different classifiers along with the hybrid classifier. Recognition results obtained in our experiments, as given in Table 1 and in Table 2, show that Curvelet-based feature and Geometry-based feature with dynamic hybrid classifier gives better result than the other classifiers and overall accuracy of 87.3%.

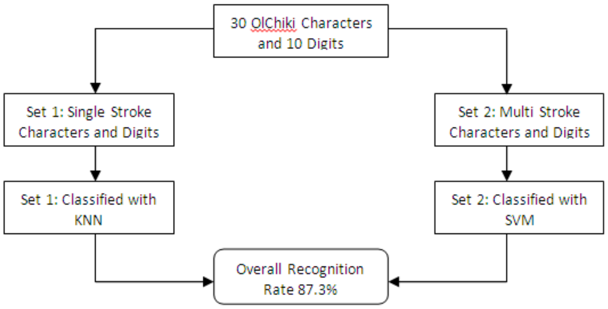

30 OlChiki Characters and 10 Digits
Set 1: Single Stroke Characters and Digits
Set 2: Multi Stroke Characters and Digits
Set 1: Classified with KNN
Set 2: Classified with SVM
Overall Recognition Rate 87.3%

Figure 4: Hybrid Classification Model

Table 1: Character and Digit recognition rate using SVM and KNN

Classifier and Features	K-NN		SVM	
	Basic Character	Digit	Basic Character	Digit
Geometry-based feature	78.4	82	74.7	75.2
Curvelet-based feature	82.9	86.1	76.3	82.3

Table 2: Character and Digit recognition rate using Hybrid Classifier

Classifier and Features	Dynamic Hybrid Classifier		
	Basic Character	Digit	Overall Accuracy
Geometry-based feature	83.9	85.6	**87.3**
Curvelet-based feature	83.87	89	

V. CONCLUSION

This paper proposes a framework for evaluation and comparison of performances of Curvelet and geometry-based hybrid features for handwritten OlChiki character recognition using k-NN and SVM classifiers dynamically and provides a better result than previous as with an accuracy level of 87%.Individually, K-NN classifier performs much better with an accuracy of 83% nearly for OlChiki basic characters recognition and 86% accuracy with OlChiki digits recognition. The novelty of this work also lies in using PCA for reducing dimensionality of Curvelet features. On the basis of our observations, Curvelet-based feature is found to be highly effective while character geometry features are not found suitable for OlChiki characters having very complex structural properties. Comparison results show that the Curvelet features are more suitable for OlChiki character recognition and with k-NN classifier gives better results than SVM. Hence, Curvelet transform in combination with K-NN classifier proves to be useful in OlChiki character recognition. However, the accuracy of proposed scheme may be enhanced by increasing the number of training samples and/or applying the proposed scheme at different resolution levels. Huge historical documents in OlChiki are conserved. Computerization of those documents may be needed in future. This paper may help for future research to recognize the word or paragraph in OlChiki Script.

ACKNOWLEDGEMENT

The authors express the deep sense of gratuity towards the Department of Computer Science and Engineering, University of Burdwan where the various resources are used for the work.

REFERENCES

[1] Marine Carrin, *"The impact of cultural diversity and globalization in developing a Santali peer culture in Middle India",EMIGRA Working Papers, ISSN 2013-3804*.

[2] R.C. Hansdah, N.C. Murmu, *"Encoding of OlChiki in Universal Character Set"* - ISO/IEC 10646.

[3] N. Otsu, *"A Threshold Selection Method from Grey-Level Histogram",* IEEE Trans Systems, Man & Cybernetics, vol. 9, no. 1, pp. 62 – 66, 1979.

[4] V.N. Vapnik, "*The Nature of Statistical Learning Theory"*, 2nd ed., Springer, 2000.

[5] Y. Yang, *"Expert Network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval"*, Proc. 17th Annual Intl. ACM SIGIR Conf. Research & Development in Information Retrieval , Dublin (Ireland), pp. 13 – 22, 1994.

[6] Singh, Mittal, Ghosh, *"An Evaluation of Different Feature Extractors and Classifiers for Offline Handwritten Devnagari Character Recognition"* – Journal of Pattern Recognition Research 2 (2011), pp. 269-277.

[7] SumantaDaw, Abhoy Chand Mondal, *"A Font Invariant OlChiki Basic Character Recognition using Digital Curvlet Transformation"*-ICCS 2013.

Authors Profile

Mr. Sumanta Daw is an Assistant Professor in the Department of Computer Science & Engineering at Hooghly Engineering & Technology College in Hooghly. He is currently pursuing Ph.D. He was born on 20th April, 1979. He obtains his M.Sc. in Software Engineering in the year of 2003 and M.Tech. in Computer Science & Engineering in the year of 2010. He worked in industry for one year and as an academician for last 14 years. He has already published several national and international journals and conference papers in the fields of Soft Computing and Network Security.

Dr. Abhoy Chand Mondal is a Professor of Department of Computer Science, The University of Burdwan. He was born on 27th February 1964. He received his B.Sc. (Math-Hons.) degree from The University of Burdwan in 1987, M.Sc. (Math) and MCA from Jadavpur University in 1989 and 1992 respectively. He received his Ph.D. degree from Burdwan University in 2004. His research interest is in Soft Computing, Document Processing, and Web Mining etc. He has 1 year industry experience and 22 years of teaching and research experience. He has published 60 papers (No. of Journal papers are 10 (Scopus Index) and 20 other UGC listed Journal Papers). So far six students have been awarded Ph.D. degree under his guidance. Currently seven students are registered for their Ph.D. work under his supervision.