Off-Line Handwritten Odia Character Recognition Using DWT and PCA

Abhisek Sethy, Prashanta Kumar Patra and Deepak Ranjan Nayak

Abstract In this paper, we propose a new approach for Odia handwritten character recognition based on discrete wavelet transform (DWT) and principal component analysis (PCA). Statistical feature descriptors like mean, standard deviation, energy have been computed from each sub-band of the second level DWT and are served as the primary features. To find the most significant features, PCA is applied. Subsequently, back-propagation neural network (BPNN) is harnessed to perform the classification task. The proposed method is validated on a standard Odia dataset, containing 150 samples from each of the 47 categories. The simulation results offer a recognition rate of 94.8%.

Keywords Discrete wavelet transform (DWT) · Optical character recognition (OCR) · Principal component analysis (PCA)

1 Introduction

Optical character recognition (OCR) systems have received much more attention over the last decades because of its wide range of applications in bank, postal, and industries. These systems have been designed not only for printed characters but also for handwritten characters. However, recognizing the handwritten characters has become a difficult task for researchers due to the variations in writing styles of

A. Sethy (✉) · P. K. Patra
Department of Computer Science and Engineering,
College of Engineering and Techonoloy, Bhubaneswar, India
e-mail: abhisek052@gmail.com

P. K. Patra
e-mail: principalcet@cet.edu.in

D. R. Nayak
Department of Computer Science and Engineering,
National Institute of Techonoloy, Rourkela, India
e-mail: depakranjannayak@gmail.com

K. Saeed et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 563,
https://doi.org/10.1007/978-981-10-6872-0_18

human beings and remains an open problem. In recent years, many OCR systems have been proposed by diverse researchers over different languages like Japanese, Chinese, and Arabic [1]. This paper aims at developing an OCR system based on handwritten Odia characters. Odia is the one of ancient and famous regional language in the eastern India and mostly spoken in the state of Odisha, and Kolkata. Recognition of Odia characters has become a tiresome and demanding task. In handwritten Odia characters, a lot of ambiguities can occur as most of the characters are similar in shape and structure and have same orientations. Therefore, it is required to design a robust OCR system that can correctly discriminate the characters. Odia language consists of 49 characters (14 vowels and 35 consonants) and some special conjunct consonants. During past years, different authors have made an attempt for analysis on Odia scripts [2]. Feature extraction stage plays an important role for better recognition. In this paper, we have applied discrete wavelet transform (DWT) on the character images for feature extraction. The coefficients of the level-2 decomposition are computed, and then statistical features like mean, median, min, max, standard deviation, mean absolute deviation, median absolute deviation, Euclidean distance and energy have been calculated for each sub-band. These values are considered as the key feature values for each character image. Thereafter, principal component analysis (PCA) has been employed to obtain the more important features, and subsequently, BPNN is utilized to classify the characters. Simulation results on a standard dataset offer 94.8% accuracy.

The remainder of this paper is structured as follows. We summarize the related works in Sect. 2. We present different methodologies adopted in the proposed system in Sect. 3. We report the simulation results in Sect. 4, and at last we draw conclusions and outline the future scope in Sect. 5.

2 Related Works

In the last two decades, numerous works have been introduced over printed and handwritten Odia scripts. Orientation, angular rotation of character images are reported by Patra et al. in [3]. The authors have used Fourier-Modified Direct Mellin Transform (FMDMT) and Zernike moments over variant and invariant scaled character to get the features. Their experiments were conducted on 49 Odia character images, and the classification accuracy of 99.98% is achieved through probabilistic neural network classifier. In [4], Pal et al. have developed an offline Odia handwritten character recognition system based on curvature feature. PCA was used to reduce 1176 dimensional feature vector to 392, and finally they got 94.6% accuracy by using modified quadratic classifier [5]. Chaudhuri et al. [6] suggested an OCR system for printed Odia script where each character is recognized by a combination of stroke, run-number based feature. In addition, water reservoir-based features were also used. Later on, a binary stage recognition system for handwritten characters is introduced by Padhi and Senapati [7]. They have calculated average zone centroid distance and also reported the mismatch among several characters for achieving high recognition

rate. They used two artificial neural networks (ANNs) for the characters of similar groups and for each individual one. Wakabayashi et al. [8] have given a comparative analysis of similar shaped characters of Odia with respect to other languages like Arabic/Persian, Devnagari, English, Bangla, Tamil, Kannada, Telugu. They have introduced Fisher ratio (F-ratio) approach for character along with gradient feature for feature extraction. Dash et al. [9] have utilized PCA for dimensional reduction for feature vector. The unconstrained numerals were taken for classification. A new method called Kirsch edge operator is used which is an edge detection algorithm. They have implemented Modified Quadratic Discriminate Function (MQDF) and Discriminative Learning Quadratic Discriminate Function (DLQDF) as the classifier and achieved 98 and 98.5% recognition rate using MQDF and DLQDF, respectively. Various authors independent writing is classified by Chand et al. in [10]. They have used support vector machine and got promising results. Kumar et al. in [11] have proposed the ant-miner algorithm (AMA) for offline handwritten Oriya character recognition. It is an extension of ant colony optimization algorithm. They have taken the matrix space analysis method and feature analysis method for analyzing the handwritten images. Stroke prevention algorithms are also established by Pujari et al. in [12]. They have proposed a new parallel thinning algorithm to preserve significant features.

The literature study reveals that works in this area are still limited. Most of them are implemented on a smaller dataset. Hence, there is a scope to enhance the recognition rate further on a larger database.

3 Proposed Method and Materials

In this section, we portray the proposed system which aims at recognizing the Odia characters and present the materials used on it. The general steps of the proposed approach are depicted in Fig. 1. which mainly consists of the following steps: We first preprocess the input data and then use DWT to extract features; we use PCA for feature reduction, and subsequently we employ BPNN for classification.

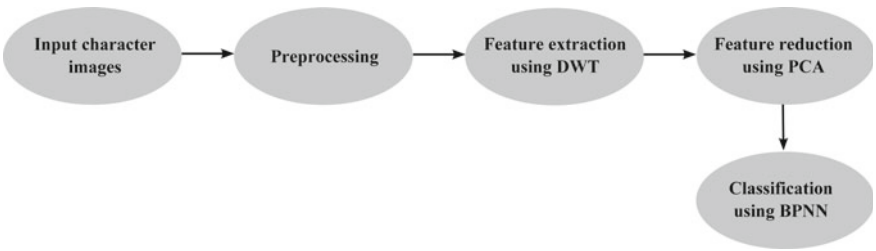

Fig. 1 Proposed model for Odia character recognition

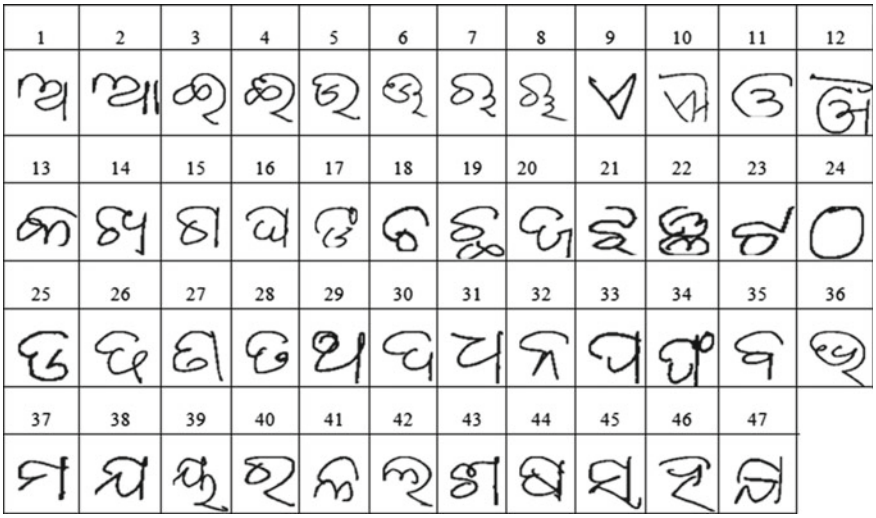

Fig. 2 Samples of 47 Odia characters

3.1 Materials

A standard database, called NIT Rourkela Odia database, has been used to validate the proposed system, which was designed by Mishra et al. in [13]. The database consists of handwritten Odia characters and numerals of different users. Around 15040 images were reported there. But in this paper, we use only 47 Odia characters which are numbered from 1 to 47 as shown in Fig. 2. We have chosen 150 numbers of samples from each character, and hence a total of $47 \times 150(=7050)$ images have been considered for simulation.

3.2 Preprocessing

Preprocessing is one of the most essential steps of any recognition system. It helps to maintain the originality of the character image by removing the unwanted things from the image. Here, we first resize the input grayscale image into 81×81. Additionally, we employ min-max normalization on the character matrix of the respective images in order to obtain a normalized data set; finally, a morphological operation called dilation is utilized.

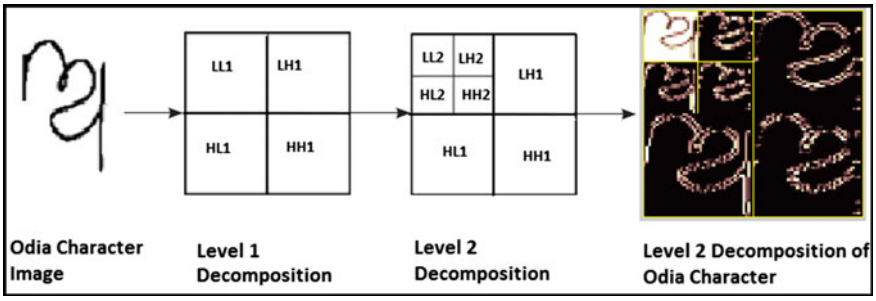

Fig. 3 Sample Odia character and its wavelet decomposition at second level

3.3 Feature Extraction Using DWT

A good feature extractor leads to a better recognition rate [14]. This paper utilizes the coefficients of level-2 decomposition of DWT for feature extraction. DWT [5] is usually an extension of Fourier transformation which analyzes the signal at different scales or resolutions and, therefore, has become a popular method for feature extraction. It is designed with a low-pass and high-pass filters along with down samplers. Whenever DWT is applied to an image, it produces sub-band images like LL, LH, HL, and HH for each individual level. Among these LH, HL, HH represents detail components in the horizontal, vertical, diagonal directions and LL are the approximation component which is further used in the second level of decomposition [15]. In this work, we have considered Haar wavelet which is orthogonal in nature. It helps to achieve high recognition rates.

Figure 3 depicts wavelet coefficients of a sample Odia character and its level-2 decomposition. This work considers the coefficient of all sub-bands at level-2 decomposition to extract the feature. The key features like mean, median, min, max, standard deviation, mean absolute deviation, median absolute deviation, Euclidean distance and energy have been computed from all the sub-bands. Therefore, we get a feature vector of length 63 (9 features from 7 sub-bands) for each character image. Eventually, a feature matrix of size 7050×63 is constructed. The obtained feature matrix is then passed to PCA to reduce the dimension further.

3.4 Feature Reduction Using PCA

The presence of insignificant features leads to high computation overhead, more storage memory and sometimes reduce the performance of the classifier. Hence, it is necessary to find the most significant features from the original feature set. In this work, we have used PCA to reduce the dimension of the feature. PCA is the most well-known approach that has been broadly harnessed for dimensionality reduction

and data visualization in many applications [16]. It projects the input data onto a lower dimensional linear space, termed as the principal subspace with an aim of maximizing the variance of the projected data. The main motivation of using PCA in this paper is to lessen the dimensionality of the features which results in a more accurate and efficient classifier.

3.5 Classification Using BPNN

Neural networks have gained popularity in classification problems because of its several advantages over probabilistic-based classifiers. It can be defined as a massively parallel processor which can learn through examples [17]. In this paper, we have used a back-propagation neural network having sigmoid activation in the hidden layer and linear activation in the output layer, to classify characters into a set of target categories. For training, we used scale conjugate gradient technique as it produces faster convergence than gradient descent approach. The input layer of BPNN consists of 9 neurons as nine features are selected by the PCA. It may be noted that the number of neurons in hidden and output layer is set to 25 and 47, respectively. The performance of BPNN was measured by mean square error (MSE) and is defined as

$$MSE = \frac{1}{n}\sum_{n}(T - O)^2 \tag{1}$$

where T is the target output, O is the actual output, and n is the total number of training data.

3.6 Implementation

The overall pseudocode of the proposed system is presented in Algorithm 1. It is divided into two phases: offline and online phase. In offline phase, we train the network based on the features extracted in the extraction and reduction step; however in online phase, we can predict a class label for an unknown sample.

4 Simulation Results and Discussions

All the methods of the proposed OCR system were simulated using MATLAB 2014a. After preprocessing, in feature extraction step we got a feature vector of length 63 for each image. Then, PCA is used to reduce the dimension to 9 as 9 principal components (PCs) can able to preserve more than 80% of the total variance. In addition, it has been found that with only nine PCs, the system earns highest accuracy. Then,

Table 1 Recognition rate achieved using the proposed system

Sl no	N_{cc}	N_{mc}	Recognition rate (%)	Sl No	$\mathbf{N_{cc}}$	N_{mc}	Recognition rate (%)
1	145	5	96.6	25	144	6	96.0
2	142	8	94.6	26	146	4	97.3
3	142	8	94.6	27	140	10	93.3
4	140	10	93.3	28	141	9	94
5	140	10	93.3	29	143	7	95.3
6	143	7	95.3	30	144	6	96.0
7	141	9	94.0	31	144	6	96.0
8	141	9	94.0	32	144	6	96.0
9	141	9	94.0	33	140	10	93.3
10	141	9	94.0	34	140	10	93.3
11	139	11	92.6	35	140	10	93.3
12	140	10	93.3	36	141	9	94.0
13	143	7	95.3	37	142	8	94.6
14	143	7	95.3	38	143	7	95.3
15	143	7	95.3	39	142	8	94.6
16	144	6	96.0	40	142	8	94.6
17	146	4	97.3	41	142	8	94.6
18	146	4	97.3	42	142	8	94.6
19	142	8	97.3	43	142	8	94.6
20	146	4	97.3	44	142	8	94.6
21	140	10	93.3	45	142	8	94.6
22	141	9	94.0	46	144	6	96.0
23	142	8	94.6	47	144	6	96.0
24	143	7	95.3	Overall recognition rate = 94.8%			

Algorithm 1 Pesudocode of the proposed system

Offline learning:

1: The input images are pre-processed and are decomposed by DWT.
2: Calculate the statistical features from all the sub-bands of 2^{nd} level DWT.
3: PCA is carried out and principal component (PC) coefficient matrix is generated.
4: The reduced set of features along with its corresponding class labels are used to train the BPNN classifier.
5: Report the performance.

Online prediction:

1: Users presented a query image to be classified
2: DWT is performed on the query image and then statistical features are calculated from all seven the sub-bands
3: PC score is obtained by multiplying feature vector into PC coefficient matrix
4: The PC score is given input to the previously trained BPNN to predict the class label

a reduced feature matrix is obtained. The input dataset is divided into 70% training and 30% testing samples. Thereafter, we design a BPNN network $9 \times 25 \times 47$ to perform classification. The overall classification results of the proposed system are listed in Table 1, where the 47 characters are numbered from 1 to 47. $\mathbf{N_{cc}}$ denotes the number of times the characters correctly classified, and $\mathbf{N_{mc}}$ indicates the number of times the characters are miss-classified. From the table, it has been observed that the overall recognition rate is 94.8%.

5 Conclusion and Future Scope

This paper presents an automatic OCR system for Odia characters and achieved 94.8% accuracy on a benchmark dataset. In the preprocessing step, we perform operations like normalization and dilation. Two-dimensional DWT has been used for feature extraction from character images followed by PCA for feature reduction. Finally, the reduce set of features is fed to the BPNN classier. For feature extraction, DWT is decomposed up to two levels; however, features from high levels of decompositions may be considered. Other feature selection techniques like filter-based techniques or evolutionary-based approaches can be applied to find the most significant features. Different combination of classifiers can also be taken into account for character recognition in the future.

References

1. Plamondon, R., Srihari, S.N.: On-line and off-line handwritten recognition: a comprehensive survey. IEEE Trans. PAMI **22**, 62–84 (2000)
2. Pal, U., Jayadevan, R., Sharma, N.: Handwriting recognition in indian regional scripts: a survey of offline techniques. ACM Trans. Asian Lang. Inf. Process. **11**(1), 1–35 (2012)
3. Patra, P.K., Nayak, M., Nayak, S.K., Gabbak, N.K.: Probabilistic neural network for pattern classification. In: International Joint Conference on Neural Networks, pp. 1200–1205 (2002)
4. Pal, U., Wakabayashi, T., Kimura, F.: A system for off-line oriya handwritten character recognition using curvature feature. In: 10th International Conference on Information Technology, pp. 227–229 (2005)
5. Pratt, W.K.: Digital Image Processing. Willey (2007)
6. Chaudhuri, B.B., Pal, U., Mitra, M.: Automatic recognition of printed oriya script. In: Sixth International Conference on Document Analysis and Recognition, pp. 795–799 (2001)
7. Padhi, D., Senapati, D.: Zone centroid distance and standard deviation based feature matrix for odia handwritten character recognition. In: International Conference on Frontiers of Intelligent Computing Theory and Applications (FICTA), vol. 199, pp. 649–658 (2005)
8. Wakabayashi, T., Pal, U., Kimura, F., Miyake, Y.: F-ratio based weighted feature extraction for similar shape character recognition. In: 10th International Conference on Document Analysis and Recognition (2009)
9. Dash, S.K., Puhan, N.B., Panda, G.: A hybrid feature and discriminate classifier for high accuracy handwritten odia numeral recognition. IEEE Region 10 Symposium, pp. 531–535 (2014)
10. Chand, S., Frank, K., Pal, U.: Text independent writer identification for oriya script. In: IAPR International Workshop on Document Analysis Systems (2012)

11. Kumar, B., Kumar, N., Palai, C., Jena, P.K., Chattopadhyay, S.: Optical character recognition using ant miner algorithm: a case study on oriya character recognition. Int. J. Comput. Appl. **61**(3), 0975–8887 (2013)
12. Pujari, A.K., Mitra, C., Mishra, S.: A new parallel thinning algorithm with stroke correction for odia characters. In: Advanced Computing, Networking and Informatics, vol. 1, pp. 413–419. Springer (2014)
13. Mishra, T.K., Majhi, B., Sa, P.K., Panda, S.: Model based odia numeral recognition using fuzzy aggregated features. Front. Comput. Sci. Springer, pp. 916–922 (2014)
14. Kumar, G., Bhatia, P.K.: A Detailed review of feature extraction in image processing systems. In: Fourth International Conference on Advanced Computing & Communication Technologies (2014)
15. Nayak, D.R., Dash, R., Majhi, B.: Brain MR image classification using two-dimensional discrete transform and adaboost with random forests. Neurocomputing **177**, 188–197 (2016)
16. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2006)
17. Haykin, S.: Neural Networks: A Comprehensive Foundation. Prentice Hall (1999)