International Journal of Electrical and Computer Engineering (IJECE)
Vol. 12, No. 6, December 2022, pp. 6103~6113
ISSN: 2088-8708, DOI: 10.11591/ijece.v12i6.pp6103-6113

Recognition of compound characters in Kannada language

Sridevi Tumkur Narasimhaiah, Lalitha Rangarajan
Department of studies in Computer Science, University of Mysore, Mysuru, Karnataka, India

Article Info

Article history:

Received Oct 19, 2021
Revised Jun 11, 2022
Accepted Jul 7, 2022

Keywords:

Deep convolutional neural networks classifier
Degraded character recognition
Histogram of oriented gradients
Old Kannada documents
Optical character recognition
Principal component analysis
dimensionality reduction

ABSTRACT

Recognition of degraded printed compound Kannada characters is a challenging research problem. It has been verified experimentally that noise removal is an essential preprocessing step. Proposed are two methods for degraded Kannada character recognition problem. Method 1 is conventionally used histogram of oriented gradients (HOG) feature extraction for character recognition problem. Extracted features are transformed and reduced using principal component analysis (PCA) and classification performed. Various classifiers are experimented with. Simple compound character classification is satisfactory (more than 98% accuracy) with this method. However, the method does not perform well on other two compound types. Method 2 is deep convolutional neural networks (CNN) model for classification. This outperforms HOG features and classification. The highest classification accuracy is found as 98.8% for simple compound character classification. The performance of deep CNN is far better for other two compound types. Deep CNN turns out to better for pooled character classes.

Corresponding Author:

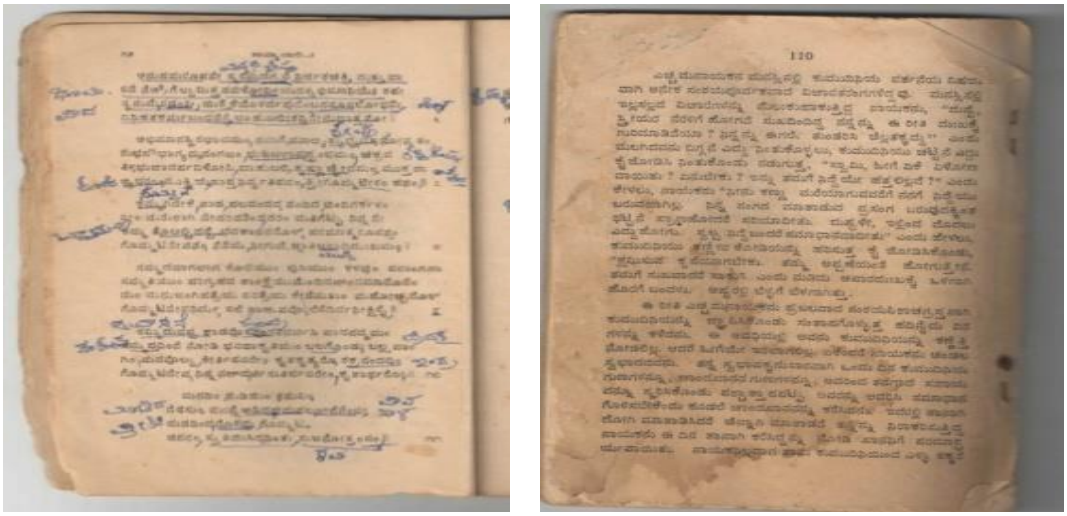
Sridevi Tumkur Narasimhaiah
Department of Studies in Computer Science, University of Mysore
Mysuru, Karnataka, India
Email: tn.sridevi1@gmail.com

1. INTRODUCTION

Recognition of characters in South Indian Script is one of the challenging research problems due to large number of classes with complex geometrical shapes of characters. Some challenges in document image recognition are complex layouts, degraded texts, characters with varying font sizes and styles, different illumination levels in different parts, artifacts created at the time of imaging, and variety of degradation factors. These problems have led to the development of algorithms that suit a specific set of problems present in the document. Recognition of compound characters in degraded Kannada printed documents is the focus of the current investigation. Segmentation of compound characters into components in degraded Kannada printed documents is a complex problem as most of the components in the compound character are broken or merged. In this research, an attempt is made to recognize the degraded compound characters which can essentially lead to enhancement of accuracy of such characters in degraded documents.

Optical character recognition (OCR) algorithms work well on non-degraded documents but not necessarily on non-degraded documents. Presence of degradations like aging marks, dilated characters, bulge in specific portions of characters, merges and splits within a character and ink marks due to use of annotations, are inevitable in old printed books or documents. Currently, OCRs are developed to address these problems and a very few address the design needs of degraded Kannada documents. Applying the existing OCR's which are designed for non-degraded documents produces poor recognition accuracies. Therefore, there exists a scope for a good methodology for degraded character recognition.

The proposed method works by extracting characters and creating a dataset of same class containing variety of degradations from documents preprocessed using binary image analysis technique (BIA) [1]. The types of degradations covered in the dataset include varying contour deformations, dilations/breakages in different portions of character geometry and distortions caused due to aging noise, and stains. The datasets of various degradation types are captured from documents belonging to different ancient books. The dataset devised covers as much possible degradations of characters belonging to same class. Figure 1 shows sample degraded documents, Figure 1(a) document with annotations Figure 1(b) incorrect illumination in the document. Figure 2 shows sample of characters extracted, Figure 2(a) before preprocessing and Figure 2(b) after preprocessing. And Figure 3 shows few samples of severely degraded compound characters subsequent to preprocessing. These characters include Figure 3(a) simple compound, Figure 3(b) multi compound and Figure 3(c) complex compound.

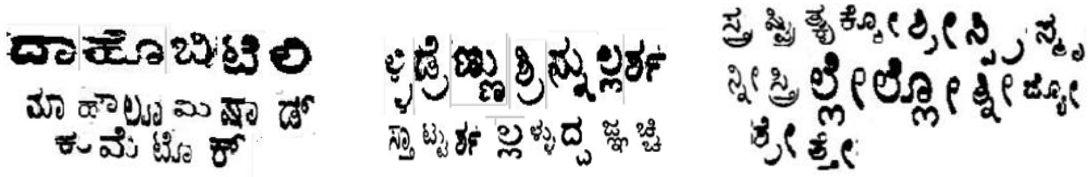


(a) (b)

Figure 1. Sample degraded documents (a) document with annotations and (b) incorrect illumination in the document

(a) (b)

Figure 2. Sample of characters extracted (a) before preprocessing and (b) after preprocessing



(a) (b) (c)

Figure 3. Compound characters (a) simple compound, (b) multi compound, and (c) complex compound

Kamble and Hegadi [2] extracted histogram of oriented gradients (HOG) features from a rectangular block suitable for real time applications. Model experiments on 8,000 samples of handwritten Marathi characters normalized to 20×20 pixel size. Classification is performed using g support vector machine (SVM) classifier and feed-forward artificial neural network (ANN). Results demonstrated high performance when classified using feed-forward ANN.

A system is proposed by Naiemi *et al.* [3] to recognize character using enhanced HOG to detect spam image. Spam image is basically an unsolicited message which is electronically sent to a larger group of arbitrary addresses. OCR is one of the ways to come across such spam images. Authors have proposed spam detection using HOG feature extraction and SVM for classification method. Model devised two heuristic procedures for thickening and thinning of characters in the pre-processing stage to improve the recognition accuracy. The average accuracy on the modified HOG method is found to be 91.6% over char74K database. Proposed method when examined on ICDAR2003 database outperforms enhancement in comparison with the previous methods.

Narang *et al.* [4] presented a work in recognition of Devanagari script from ancient documents. discrete cosine transformation (DCT) zigzag features and HOG are used for extracting the features. Classification is performed using SVM, decision tree and naïve Bayes. They created a database from various libraries and museums. Highest recognition accuracy of 90.70% has been achieved using SVM classifier with zigzag feature vector of length 100. Model considered training and testing ratio of 80% and 20% respectively. A new genetic algorithm based feature selection approach is proposed by Cordella *et al.* [5]. Model aims at selecting the smallest subset which optimizes the class separation. Algorithm experimented on handwritten digits for both feature reduction has shown improvement of classifier performance. Huang *et al.* [6] describe a recognition system for handwritten offline Pashto characters. A database of 11,352-character images is considered for experimentation. Features are extracted using HOG and zoning-based density features. K-nearest neighbor classifier results in an accuracy of 80.34% for HOG and 76.42% for zoning-based density features using 10-fold cross validation. An attempt is made by Reta *et al*. [7] to address the challenges and difficulties of Amharic handwritten character recognition. They combined various feature extraction techniques such as HOG, local binary pattern (LBP) and geometrical features. Combined features are reduced using linear discriminant analysis (LDA). Classification is performed using multiclass SVM with error correcting output code (ECOC) framework. 74K benchmark numeric data sets have been considered for experimentation. Performance of the model is evaluated using 10-fold cross validation technique. Combined feature extraction and multiclass SVM classification has resulted in good recognition.

A novel feature extraction technique is proposed by Soora and Deshpande [8] towards multilingual character recognition of Indian scripts namely, English, Devanagari, and Marathi. They present a set of feature vectors (FVs) based on shape geometry. The first FV uses the triangular area to decode the input character. Second FV is extracted by dividing image into individual components. Further each component is decoded into shape symbols by comparing perpendicular distances of the individual pixels of the components. These distances are normalized. An appreciable performance has been achieved by conducting the experiments on media-lab license plate benchmark database. Proposed is a robust recognition of characters by Newell and Griffin [9]. Model considers two extensions of HOG descriptors to have features at multiple scales. Performance has been evaluated using characters extracted from images and graphics. Combined pairs of HOG at different scales achieve an improved accuracy of 12.4% and 5.6% on two datasets. Singh *et al.* [10] presented a method for recognition of handwritten Devanagari numerals. Work is focused on the feature selection based upon information theory measures. Classification is performed using multilayer perceptron (MLP) based classifier combination with feature selection using functions namely, i) maximum relevance minimum redundancy and ii) conditional mutual information maximization. Most of the benchmark datasets are considered for experimentation. Conditional mutual information-based feature selection used with the ensemble of classifier results in improved recognition. Method gives quite efficient results using only 10% as a training set. Kulkarni [11] employs k nearest neighbor (KNN) algorithm for OCR of patterns. Accuracy achieved is 90%. It is observed that the model is computationally expensive and not suitable for large dataset. The model uses HOG feature extraction with SVM algorithm for center of mass of image. Model experiments on 168233 samples of 369 classes. Accuracy of the model is found to be 96.56%.

Su *et al.* [12] made an attempt towards recognizing the characters in natural languages present in complex background with variations in text size. A new technique has been proposed with the use of convolutional cooccurrence histogram of oriented gradient (ConvCoHOG) which is more robust and differential than both HOG and co-occurrence histogram of oriented gradients (CoHOG). Informative features are constructed by extracting features from image patches. Experiments are conducted on two datasets, ICDAR 2003 and street view text (SVT) dataset. Proposed model achieves superior performance than state-of-the art techniques. A new method by Zhang *et al.* [13] has been proposed on hybrid feature extraction and selection for improving the accuracy of handwritten numerals. Following seven feature extraction methods are used: i) gradient-based wavelet features, ii) medial axial transformation algorithm

(MAT) based directional features, iii) complex wavelet features, iv) binary gradient directional features, v) median filter gradient features, vi) image thinning distance features, and vii) geometrical features. Modified National Institute of Standards and Technology (MNIST) database has been considered for experimentation and the proposed model shows improved recognition. Offline handwritten recognition system is discussed in Sethy and Patra [14] on Odia characters. They have used R-HOG for feature extraction and principal component analysis (PCA) for feature reduction. Linear SVM classifier performs better than quadratic with the recognition accuracies 98.8% and 96.8% respectively. 10-fold cross validation has been used for reliable performance estimation. New method has been proposed by Bag *et al.* [15] on Bangla compound characters. Model extracts complex shape primitives and uses a template matching to recognize the characters. Novelty of their work lies in segmenting character skeleton into stroke segments and merging them to extract meaningful shape components. Model is tested on both printed and handwritten characters. Proposed method works better for complex-shaped compound characters.

Pal *et al*. [16] in their work have expressed difficulty in recognition of Indian script due to the presence of complex shaped compound (cluster) characters. Modified quadratic discriminant function (MQDF) has been used to recognize off-line Bangla handwritten compound characters. Features are mainly based on arc tangent of the gradient by applying the 2×2 mean filters 4 times and Roberts filter on the gray level image along with non-linear size normalization. Finally, frequencies of directions are down sampled with the use of Gaussian filter to get 392-dimensional feature vectors. They used 5-fold cross validation technique and obtained 85.90% accuracy on 20,543 samples. Sauvola and Pietikäinen [17] proposed a new approach on binarizing a document image. They considered a hybrid approach and document region class properties. Local contents of a page to the background, pictures, and text have been quickly classified by the model. Model implements two different approaches to define a threshold for each pixel: i) soft decision method (SDM) for pictures and background and ii) text binarization method (TBM) for isolating the textual and line areas from poorly illuminated background. Output of these two algorithms has been combined finally. The model is evaluated using ground truth test images. It uses evaluation metrics for binarizing textual and synthetic images. The results are compared with other models in literature and authors claim that the model works well both quantitatively and qualitatively. An intensive survey on various thresholding methods has been carried out by Sahoo *et al.* [18]. The model notices the enhancement regions where the thresholding algorithms fail to modify background intensity dominates over the foreground regions. The model evaluates some automatic global thresholding methods. Here the parameters used are uniformity and shape measures.

An attempt is made to restore images from ancient post card by Roe and Mello [19]. Model uses different color image enhancement techniques which include Gaussian filtering difference, edge detection, background segmentation, noise spot detection, histogram equalization techniques. The work considers the illumination problem. They used the color constancy check. The postcards considered for experimentation are from nineteenth century. Model achieves satisfactory visual results. An attempt is made by Bannigidad and Gudada [20] to identify age-type (script of the dynasties) and recognize historical Kannada handwritten documents. Model applied image enhancement techniques to reconstruct, digitize, and recognize the document images. Features are extracted using HOG feature descriptors. KNN and SVM classifiers are used for recognition. The average recognition accuracy obtained for different dynasties are observed as 92.3% and 96.7% for KNN and SVM classifiers respectively. A robust system for printed and handwritten character recognition has been proposed by Bahi *et al.* [21] on the images obtained from camera phone. Initially model discusses the problems in building an efficient OCR, and the problems encountered with the blurred and noisy image. Model focuses on collecting the images through camera phone. The different phases in the model involve preprocessing, segmentation, feature extraction and classification. They investigated different techniques in preprocessing before choosing the best one suited for their datasets. Similar investigations have been carried out in choosing feature extraction and classification algorithms. Finally, recognition has been carried out using SVM, Naïve Bayes and multilayer network.

Saini *et al.* [22] proposed a framework known as KannadaRes-NeXt to recognize Kannada numerals. Model classifies the images using Res-NeXt. Kannada-MNIST uses 60,000 training images and 10,000 testing images. Dig-MNIST dataset contain a test sample of 10,240 images. Two more sets called AugKannada-MNIST and AugDig-MNIST contains 10,000 and 10,240 samples respectively. An accuracy of 97.36% achieved with Kannada-MNIST and 79.06% with Dig-MNIST respectively. A deep convolution network is deployed by Chandrakala and Thippeswamy [23] to recognize historical Kannada handwritten characters. Feature extraction and classification is unified by the model. Features are extracted from the characters using deep convolutional neural networks (DCNN). Features extracted are recognized using stochastic gradient descent with momentum (SGDM) and SVM algorithms. Digitized estampages are datasets considered for experimentation. An accuracy of 70% is observed with Alexnet. Hallur and Hegadi [24] classified a handwritten Kannada numerals using deep CNN. Here the handwritten Kannada numerals

are given in a document mode. Initially document is subjected to the preprocessing steps that include noise elimination, binarization, normalization, thinning, and skew correction. Discrete wavelet transform (DWT), curvelet transfiguration wrapping, drift length count and direction related progression code has been used for feature extraction. Recognition is performed using deep CNN and observed accuracy is 96%.

It is evident from literature that various attempts are reported on recognition of degraded characters from various Indian and non-Indian scripts. Work focused mostly on the feature extraction techniques on printed/handwritten numerals and characters. Very few works have been reported on compound characters in Indian languages. Not many of the works deal on compound characters from non-degraded/degraded printed Kannada documents. Hence the proposed method focuses on feature extraction and recognition of compound characters particularly from degraded documents. Also, it is necessary to address variety of degradations associated with compound characters particularly found in old document images.

The rest of the paper is organized in three sections. Sections 2 covers character classifications on raw image, section 3 covers proposed methodology consisting of database creation, feature extraction using HOG followed by PCA. In section 4 results of experimental analysis on HOG features followed by classification and deep CNN for recognition are discussed. Section 5 concludes the paper.

2. CHARACTER CLASSIFICATIONS ON RAW IMAGE

Experiments are conducted on retrieved characters from pages of degraded documents. Sample of 50 characters in each category (simple, multi, complex compound) is chosen and different orientations of these characters are stored in the database. Database characters cover various types of degradations in the character set. Table 1 gives details of data set.

Table 1. Dataset of unprocessed characters

Datasets	Compound characters	No. of classes	Total characters
Raw data with noise	Simple compound	50	200
	Multi compound	50	200
	Complex compound	50	272

It may be observed that in literature, HOG features are used commonly for character recognition. Experiments have been conducted using HOG feature extraction on 150 classes of various compound characters. HOG features are extracted and features are reduced using PCA (85% variation covered) and different classifiers are used. The results of experimentation may be found in Tables 2 to 4. 4×4 and 8×8 block sizes are used for extracting HOG features. More details of HOG features may be found in next section. Observe that on all types of characters all the classifiers perform poorly. Thus, pre-processing is an inevitable first step for character recognition, particularly when the characters are from degraded document pages. The result of classification accuracies of raw data of multi compound characters are shown in Table 3. Table 4 shows the result of classification accuracies of raw data of complex compound characters. Table 2 shows the result of classification accuracies of raw data of simple compound characters.

Table 2. Classification accuracies of simple compound characters on raw data

Feature	No. of features HOG/PCA	Validation	Fine Gaussian SVM	Fine KNN (k=1)	Medium KNN (k=10)	WKNN (k=10)	Cosine KNN (k=10)	Cubic KNN (k=10)	Ensemble Adaboost Classifier
HOG(4×4)	196/112	3-fold	3.0%	5.6%	4.11%	6.6%	8.1%	4.6%	10.7%
	196/67	4-fold	7.6%	8.6%	4.6%	8.1%	9.1%	4.6%	20.8%
	196/67	5-fold	0.5%	8.6%	4.1%	7.1%	11.2%	4.6%	15.2%
HOG(8×8)	144/17	3-fold	2.5%	4.1%	3.6%	4.6%	4.1%	5.1%	8.1%
	144/17	4-fold	4.6%	6.1%	4.1%	5.6%	5.1%	5.6%	5.6%
	144/17	5-fold	1.5%	3.6%	4.1%	4.6%	5.6%	3.6%	7.6%

Table 3. Classification accuracies of multi compound characters on raw data

Feature	No. of features HOG/PCA	Validation	Fine Gaussian SVM	Fine KNN (k=1)	Medium KNN (k=10)	WKNN (k=10)	Cosine KNN (k=10)	Cubic KNN (k=10)	Ensemble Adaboost Classifier
HOG(4×4)	199/75	3-fold	2.0%	2.0%	1.5%	1.5%	4.5%	3.0%	11.5%
	199/75	4-fold	2.0%	4.0%	4.5%	2.0%	3.5%	3.5%	13.0%
	199/75	5-fold	0.0%	2.5%	2.5%	3.0%	5.0%	3.5%	13.5%
HOG(8×8)	144/19	3-fold	3.0%	2.5%	3.0%	3.0%	4.0%	2.0%	4.5%
	144/19	4-fold	4.5%	2.0%	2.0%	2.0%	4.0%	3.0%	8.0%
	144/19	5-fold	0.5%	1.5%	1.0%	2.0%	3.0%	2.5%	6.5%

Table 4. Classification accuracies of complex compound characters on raw data

Feature	No. of features HOG/PCA	Validation	Fine Gaussian SVM	Fine KNN (k=1)	Medium KNN (k=10)	WKNN (k=10)	Cosine KNN (k=10)	Cubic KNN (k=10)	Ensemble Adaboost Classifier
HOG(4×4)	271/83	3-fold	21.0%	26.5%	19.9%	22.8%	21.0%	19.5%	20.6%
	271/83	4-fold	20.6%	25.0%	21.7%	23.5%	22.8%	20.6%	22.1%
	271/83	5-fold	21.0%	28.7%	19.9%	22.8%	21.3%	18.8%	21.3%
HOG(8×8)	144/20	3-fold	20.2%	21.7%	23.2%	24.6%	21.3%	22.8%	19.5%
	144/20	4-fold	19.9%	23.2%	22.1%	24.3%	23.9%	21.3%	19.5%
	144/20	5-fold	19.9%	23.9%	21.7%	25.0%	22.1%	22.8%	18.4%

3. PROPOSED METHOD

Multi and complex compound characters have more than one component. Many research attempts have been shown to be successful on such characters with the traditional connected component analysis approach. Our experiments using connected component extraction on degraded characters turned out to be failure. Hence this proposal of HOG features on all compound characters. The proposed model for feature extraction and classification of degraded Kannada characters starts with acquisition of characters extracted from training datasets pre-processed using BIA [25], followed by feature extraction, dimensionality reduction and recognition. The processing stages of the proposed approach are as shown in Figure 4. Extracted features for the proposed recognition model are HOG features for two different block sizes (4×4, 8×8) and these features are reduced and classification is performed using: Fine Gaussian support vector machines, Fine K-nearest neighbor (KNN), medium KNN, weighted KNN (WKNN), Cosine KNN, Cubic KNN, and ENSEMBle Adaboost classifiers.

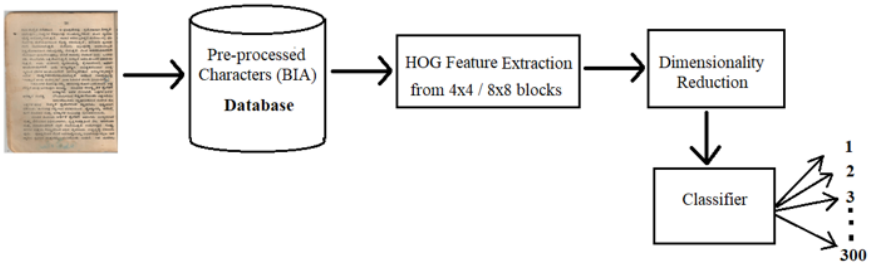


Figure 4. HOG feature extraction and classification of degraded Kannada characters

3.1. Database creation

The process of dataset generation for degraded printed character recognition is explained in this section. Higher recognition accuracy is one of the key focus in the development of OCR system. It is to be noted that classification of raw characters on degraded documents yields very poor result in Tables 2 to 4 of section 2. Hence, the documents are subjected to customized noise removal algorithm BIA [25]. Subsequently, characters are segmented using projection profile and stored in the database.

Characters from 75 pre-processed degraded pages from 10 books are extracted. In this work the characters considered are vowels, consonants, simple, multi and complex compound characters (vowel consonant clusters). Multiple instances of the extracted characters are represented in 300 classes. As the character collection is from different sources, age and storage methods vary, the collection covers different degradations and hence intra class variation is large. Different orientations of each degraded character sample are included in the dataset so that OCR can recognize documents in any direction.

It is required to build the database for training by including characters with different types of degradations, so that classification accuracy is better. Figure 3 shows some samples of simple, multi, and complex compound characters extracted from preprocessed degraded documents. Database consists of 7,360, 704, 400 simple, multi, and complex compound characters distributed in 100 classes in each category. Each preprocessed character is normalized to 28×28 and stored in database.

3.2. HOG feature extraction and dimensionality reduction

Proposed model uses the widely used HOG feature descriptor as HOG feature descriptors precisely describe the features in local cells. These descriptors are highly advantageous due to the shape sensitive nature of characters in Kannada language. In case of printed Kannada characters, the orientations present at localized regions of image remains static, and hence HOG features are apt to describe the character instances.

HOG features are computed for each character image of size 28×28, by dividing it into non overlapping blocks. Making blocks of images reduces the number of features significantly. A block (cell) is considered as a pixel grid in which gradients are computed using the magnitude and the direction of change in the intensities of the pixel within a block. For instance, from one such grid of size 4×4, a block of 16 pixels, horizontal and vertical gradients are calculated followed by gradient magnitude and gradient angle for each of 16 pixels. These computed gradient angles are distributed into 9 bins (feature values) (0-20, 20-40, … 160-180). Subsequently the features of these smaller blocks are concatenated to 36 features of a block of 4 times the size (8×8). These bigger blocks are overlapping and there are 36 such overlapping blocks. Thus, we have a total of 36x36 (1296) features describing a character. Similarly, if the initial block size is 8×8 the total number of features will be 144.

Table 5 gives details of number of HOG features for different block sizes. Number of features is too many for a 2×2 block. Many features of 2×2 blocks of distinct characters can result in similar feature values. 16×16 blocks are too big and the features extracted are likely to be far less meaningful and distinguishing. Hence, the experimentation is carried out with HOG features extracted for sizes 4×4 and 8×8.

Table 5. HOG feature extraction on various block (cell) sizes

Block (cell) size	Total features extracted
2×2	6084
4×4	1296
8×8	144
16×16	36

Principal component analysis (PCA) is employed to further reduce the features extracted using HOG. Use of PCA helps data representation in a better form with the minimal loss of information. PCA also ensures more effective data analysis on the reduced dimensional space. The resultant feature vectors are submitted to the classifier for recognition of character class.

4. EXPERIMENTAL RESULTS

The number of occurrences of vowels, consonants, simple compound characters varies extensively. In other words, we have class imbalance problem. Literature suggests ways of handling such problems. Some of these are: up sampling of minority classes, down sampling of majority classes, identify a suitable classifier. First two of the three methods are suited for poor sampling methods. Our problem of imbalance is not due to incorrect sampling. Imbalance is embedded in the data itself. Not all characters in a language will occur with equal frequency. Hence, the right way to address the current imbalance problem is to identify the right classifier.

Note that the number of classes is too large and so is the data size. We conducted experiments using simulation with smaller data set but similar (a smaller number of classes with varying frequencies proportional to the occurrences of characters in real data) and identified the classier that can be successful for the current scenario. The experiment is to try classifiers with and without imbalance and check which one gives almost identical accuracies. Table 6 gives the details of dataset. Tables 7 and 8 are the results of k-fold validation for k=5, 10. Table 9 is the result of 'holdout' validation. Note that many classifiers are experimented with thoroughly in all cases. Observe that classifier performance is not affected by imbalance problem. Further, all classifiers perform almost identically except SVM. Fine KNN is found to be best of all classifiers. Table 6 shows the description of dataset. Table 7 shows the result of 5-fold validation. Result of 10-fold validation is shown in Table 8. Table 9 shows the result of hold-out validation.

Table 6. Description of dataset

Class labels	Characters	Class size in Database1 (1198 characters)	Class size in Database 2 (1244 characters)
1	C	100	100
2	D	100	100
3	E	100	100
4	F	100	100
5	M	100	100
6	R	4	50
7	¢	128	128
8	ಀ	134	134
9	«	136	136
10	AiÀiÁ	148	148
11	¼É	148	148

Table 7. 5-fold cross validation performance accuracy of classifiers

Database	Feature	Fine Gaussian SVM	Fine KNN (k=1)	Medium KNN (k=10)	WKNN (k=10)	Cosine KNN (k=10)	Cubic KNN (k=10)	Ensemble Adaboost Classifier
Database 1 1198	HOG(4×4)	82.4%	98.6%	98.7%	98.7%	98.7%	98.2%	96.4%
Database 2 1244		83.5%	98.9%	99.0%	99.0%	99.0%	99.4%	96.1%

Table 8. 10-fold cross validation performance accuracy of classifiers

Database	Feature	Fine Gaussian SVM	Fine KNN (k=1)	Medium KNN (k=10)	WKNN (k=10)	Cosine KNN (k=10)	Cubic KNN (k=10)	Ensemble Adaboost Classifier
Database 1 1198	HOG(4×4)	83.0%	98.5%	98.8%	98.6%	98.6%	98.9%	96.6%
Database 2 1244		82.0%	98.9%	99.0%	99.0%	99.0%	99.6%	96.1%

Table 9. 25% Hold-out validation performance accuracy of classifiers

Database	Feature	Fine Gaussian SVM	Fine KNN (k=1)	Medium KNN (k=10)	WKNN (k=10)	Cosine KNN (k=10)	Cubic KNN (k=10)	Ensemble Adaboost Classifier
Database 1 1198	HOG(4×4)	76.6%	99.0%	98.7%	98.7%	99.0%	98.3%	97.0%
Database 2 1244		79.7%	98.1%	98.1%	98.1%	98.4%	97.4%	91.0%

4.1. HOG features for recognition

We have experimented with a sample of all compound characters covering various degradations. Sample set consists of 300-character classes 100 from simple, 100 from multi and 100 from complex compound character classes. The total number of characters is 7,360, 704, 400 which include multiple degradations and orientations in each of these classes. It is possible that the best classifier for small number of classes may not be the best for a larger class set. Hence, experiments are conducted with as many classifiers as in the previous case. The features are reduced using PCA covering 85% variation.

Table 10 shows results of recognition accuracies of HOG features and various classifiers on simple compound characters. The results of recognition accuracies of HOG features and various classifiers on multi compound characters are given in Table 11. Table 12 shows results of recognition accuracies of HOG features and various classifiers on complex compound characters.

Table 10. Classification accuracies of simple compound characters

Feature	No. of features HOG/PCA	Validation	Fine KNN (k=1)	Medium KNN (k=10)	WKNN (k=10)	Cosine KNN (k=10)	Cubic KNN (k=10)	Ensemble Adaboost Classifier
HOG(4×4)	1296/84	3-fold	98.5%	92.1%	98.3%	91.9%	92.2%	41.3%
	1296/84	4-fold	98.6%	93.0%	98.4%	93.0%	92.9%	36.5%
	1296/84	5-fold	98.6%	93.5%	98.5%	93.41%	93.3%	37.3%
HOG(8×8)	144/19	3-fold	97.7%	90.3%	96.9%	90.1%	90.1%	19.9%
	144/19	4-fold	97.8%	91.4%	97.3%	91.5%	91.2%	20.2%
	144/19	5-fold	97.7%	91.8%	97.2%	91.8%	91.8%	18.6%

Table 11. Classification accuracies of multi compound characters

Feature	No. of features HOG/PCA	Validation	Fine KNN (k=1)	Medium KNN (k=10)	WKNN (k=10)	Cosine KNN (k=10)	Cubic KNN (k=10)	Ensemble Adaboost Classifier
HOG(4×4)	702/81	3-fold	27.5%	25.7%	28.7%	26.9%	25.9%	20.1%
	702/81	4-fold	28.2%	26.2%	29.6%	26.7%	25.7%	23.5%
	702/81	5-fold	29.7%	28.6%	32.3%	30.3%	27.7%	20.9%
HOG(8×8)	144/19	3-fold	28.2%	27.0%	30.0%	25.71%	26.3%	16.1%
	144/19	4-fold	28.2%	27.7%	29.4%	26.91%	27.7%	17.2%
	144/19	5-fold	28.0%	27.2%	30.4%	26.6%	28.4%	13.5%

Experiments are conducted by pooling all compound characters in one database. The tables next show the classification accuracies obtained with HOG followed by PCA. It may be observed that a better accuracy of 88.0% and 87.4% as shown in Table 13 is achieved with Fine KNN classifier with k=1 for cell

sizes 4x4 and 8x8. Surprisingly Adaboost performance is very poor, where as the result of this is good when the database is small in Table 7 to Table 9. Perhaps some classifiers in Adaboost do not suit the dataset or the sequence of classifiers is improper.

Table 12. Classification accuracies of complex compound characters

Feature	No. of features HOG/PCA	Validation	Fine KNN (k=1)	Medium KNN (k=10)	WKNN (k=10)	Cosine KNN (k=10)	Cubic KNN (k=10)	Ensemble Adaboost Classifier
HOG(4×4)	399/86	3-fold	0.5%	1.0%	0.5%	1.3%	0.8%	2.5%
	399/86	4-fold	0.0%	0.5%	0.0%	0.8%	0.0%	4.8%
	399/86	5-fold	0.3%	0.0%	0.3%	0.8%	0.3%	4.8%
HOG(8×8)	144/20	3-fold	0.5%	0.5%	0.5%	0.0%	0.5%	1.3%
	144/20	4-fold	0.3%	0.0%	0.0%	0.3%	0.3%	1.5%
	144/20	5-fold	0.5%	0.0%	0.0%	0.0%	0.3%	1.3%

Table 13. Classification accuracies of sample set on preprocessed data

Feature	No. of features HOG/PCA	Validation	Fine KNN (k=1)	Medium KNN (k=10)	WKNN (k=10)	Cosine KNN (k=10)	Cubic KNN (k=10)	Ensemble Adaboost Classifier
HOG(4×4)	1296/94	5-fold	88.0%	83.3%	88.0%	83.6%	83.0%	31.8%
HOG(8×8)	144/20	5-fold	87.4%	82.1%	87.0%	82.2%	81.7%	13.2%

Table 14 shows the accuracy of online OCR and Fine KNN. From Table 14 it is observed that the efficiency of the proposed sequence of steps outperforms the efficiency of currently available online OCR [26] when tested across random 30 samples and choosing 10 characters at a time from each class (simple, multi, complex compound characters). It is also observed that there is a steep drop in the accuracy of online OCR when degraded characters are tested and are mis-recognized as different characters and ASCII symbols by the current available online OCR [26]. Proposed HOG tested across pooling all compound characters in one database of characters yielded far more accuracy rate as compared to online OCR.

Table 14. Accuracy of online OCR

Features		Accuracy of Fine KNN (k=1)	Accuracy of online OCR with 30 samples	Average accuracy of online OCR with 10-character samples		
				Simple compound	Multi compound	Complex compound
HOG	4×4	88.0%	3.3%	20%	20%	0%

An efficient deep learning model using convolution neural networks yielded better results towards recognition of degraded printed Kannada characters collected from old Kannada documents [25]. Tables 15 to 17 are the results of deep CNN classifier with average pooling for 2×2 pool size by considering a dataset derived from the dataset above. Deep CNN is known to work well when the classes sizes are more or less are the same. Hence, for this experiment the class sizes are balanced and the size of class is set to be 50. With 300 distinct classes the size of the data set is 15,000. Experiments are conducted by pooling all compound characters in one database results in Table 18.

Table 15. Classification accuracies of simple compound characters

Train to test ratio	Accuracy	Iterations per epoch	Elapsed time (Training time)	Learning rate
30%-70%	70.40%	11	13 sec	0.01
40%-60%	84.97%	15	13 sec	0.01
50%-50%	95.16%	19	15 sec	0.01
75%-25%	97.92%	29	22 sec	0.01
90%-10%	98.80%	35	23 sec	0.01

Table 16. Classification accuracies of multi compound characters

Train to test ratio	Accuracy	Iterations per epoch	Elapsed time (Training time)	Learning rate
30%-70%	63.14%	11	1 min 1 sec	0.01
40%-60%	78.47%	15	12 sec	0.01
50%-50%	81.56%	19	14 sec	0.01
75%-25%	88.67%	29	22 sec	0.01
90%-10%	89.20%	35	22 sec	0.01

Table 17. Classification accuracies of complex compound characters

Train to test ratio	Accuracy	Iterations per epoch	Elapsed time (Training time)	Learning rate
30%-70%	75.86%	11	49 sec	0.01
40%-60%	88.00%	15	12 sec	0.01
50%-50%	91.04%	19	14 sec	0.01
75%-25%	95.42%	29	20 sec	0.01
90%-10%	96.60%	35	23 sec	0.01

Table 18. Classification accuracies of (simple+multi+complex) compound characters

Train to test ratio	Accuracy	Iterations per epoch	Elapsed time (Training time)	Learning rate
30%-70%	80.02%	35	46 sec	0.01
40%-60%	89.78%	46	48 sec	0.01
50%-50%	92.82%	58	57 sec	0.01
75%-25%	94.85%	89	1 min 8 sec	0.01
90%-10%	95.62%	105	1 min 14 sec	0.01

Table 15 shows the classification accuracies of simple compound characters. The results of classification accuracies of multi compound characters are shown in Table 16. Table 17 shows the classification accuracies of complex compound characters. The table next in Table 18 shows the classification accuracies obtained with deep CNN architectures on the same data set.

5. CONCLUSION

The proposed work addresses the recognition of compound Kannada characters. A new dataset encompassing variety of degradations and multiple instances of characters belonging to same class is devised. Initially experiments conducted on raw data led to the conclusion that data needs to be preprocessed. The data set is inherently imbalanced. Experiment on typical smaller database conducted using different classifiers to identify a suitable classifier for a class imbalance problem. Subsequently, the complete dataset is classified using HOG feature descriptors followed by dimensionality reduction. Simple compound characters are classified satisfactorily whereas the other two compound character classifications are poor. Deep CNN model is a better fit for all compound types as well as for pooled data types.

REFERENCES

[1] S. T. Narasimhaiah and L. Rangarajan, "Binary Image Analysis Technique for Preprocessing of Excessively dilated characters in Aged Kannada Document Images," *International Journal of Recent Technology and Engineering*, vol. 8, no. 4, pp. 6660–6669, Nov. 2019, doi: 10.35940/ijrte.D9101.118419.

[2] P. M. Kamble and R. S. Hegadi, "Handwritten marathi character recognition using R-HOG feature," *Procedia Computer Science*, vol. 45, pp. 266–274, 2015, doi: 10.1016/j.procs.2015.03.137.

[3] F. Naiemi, V. Ghods, and H. Khalesi, "An efficient character recognition method using enhanced HOG for spam image detection," *Soft Computing*, vol. 23, no. 22, pp. 11759–11774, Jan. 2019, doi: 10.1007/s00500-018-03728-z.

[4] S. R. Narang, M. K. Jindal, and P. Sharma, "Devanagari ancient character recognition using HOG and DCT features," in *2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, Dec. 2018, pp. 215–220, doi: 10.1109/PDGC.2018.8745903.

[5] L. Cordella, C. De Stefano, F. Fontanella, and C. Marrocco, "A feature selection algorithm for handwritten character recognition," in *19th International Conference on Pattern Recognition*, Dec. 2008, pp. 1–4, doi: 10.1109/ICPR.2008.4761834.

[6] J. Huang, I. Ul Haq, C. Dai, S. Khan, S. Nazir, and M. Imtiaz, "Isolated handwritten pashto character recognition using ak-nn classification tool based on zoning and hog feature extraction techniques," *Complexity*, vol. 2021, pp. 1–8, Mar. 2021, doi: 10.1155/2021/5558373.

[7] B. Y. Reta, D. Rana, and G. V. Bhalerao, "Amharic handwritten character recognition using combined features and support vector machine," in *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*, May 2018, pp. 265–270, doi: 10.1109/ICOEI.2018.8553947.

[8] N. R. Soora and P. S. Deshpande, "Novel geometrical shape feature extraction techniques for multilingual character recognition," *IETE Technical Review*, vol. 34, no. 6, pp. 612–621, Nov. 2017, doi: 10.1080/02564602.2016.1229583.

[9] A. J. Newell and L. D. Griffin, "Multiscale histogram of oriented gradient descriptors for robust character recognition," in *Proceedings of the International Conference on Document Analysis and Recognition*, Sep. 2011, pp. 1085–1089, doi: 10.1109/ICDAR.2011.219.

[10] P. Singh, A. Verma, and N. S. Chaudhari, "Feature selection based classifier combination approach for handwritten Devanagari numeral recognition," *Sadhana*, vol. 40, no. 6, pp. 1701–1714, Sep. 2015, doi: 10.1007/s12046-015-0419-x.

[11] R. L. Kulkarni, "Handwritten character recognition using HOG, COM by OpenCV and Python," *International Journal of Advance Research in Computer Science and Management Studies*, vol. 5, no. 4, pp. 36–40, 2017.

[12] B. Su, S. Lu, S. Tian, J. H. Lim, and C. L. Tan, "Character recognition in natural scenes using convolutional co-occurrence HOG," in *22nd International Conference on Pattern Recognition*, Aug. 2014, pp. 2926–2931, doi: 10.1109/ICPR.2014.504.

[13] P. Zhang, T. D. Bui, and C. Y. Suen, "Hybrid feature extraction and feature selection for improving recognition accuracy of handwritten numerals," in *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, 2005, vol. 1,

pp. 136–140, doi: 10.1109/ICDAR.2005.129.
[14] A. Sethy and P. K. Patra, "R-HOG feature-based off- line odia handwritten character recognition," in *Examining Fractal Image Processing and Analysis*, 2020, pp. 196–210, doi: 10.4018/978-1-7998-0066-8.ch010.
[15] S. Bag, G. Harit, and P. Bhowmick, "Recognition of Bangla compound characters using structural decomposition," *Pattern Recognition*, vol. 47, no. 3, pp. 1187–1201, Mar. 2014, doi: 10.1016/j.patcog.2013.08.026.
[16] U. Pal, T. Wakabayashi, and F. Kimura, "Handwritten Bangla compound character recognition using gradient feature," in *10th International Conference on Information Technology*, Dec. 2008, pp. 208–213, doi: 10.1109/icit.2007.62.
[17] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33, no. 2, pp. 225–236, Feb. 2000, doi: 10.1016/S0031-3203(99)00055-2.
[18] P. K. Sahoo, S. Soltani, and A. K. C. Wong, "A survey of thresholding techniques," *Computer Vision, Graphics, and Image Processing*, vol. 41, no. 2, pp. 233–260, Feb. 1988, doi: 10.1016/0734-189X(88)90022-9.
[19] E. Roe and C. A. B. de Mello, "Restoring images of ancient color postcards," *The Visual Computer*, vol. 31, no. 5, pp. 627–641, May 2015, doi: 10.1007/s00371-014-0988-4.
[20] P. Bannigidad and C. Gudada, "Age-type identification and recognition of Historical Kannada handwritten document images using HOG feature descriptors," in *Advances in Intelligent Systems and Computing*, vol. 810, Springer Singapore, 2019, pp. 1001–1010, doi: 10.1007/978-981-13-1513-8_101.
[21] H. El Bahi, Z. Mahani, A. Zatni, and S. Saoud, "A robust system for printed and handwritten character recognition of images obtained by camera phone," *WSEAS Transactions on Signal Processing*, vol. 11, pp. 9–22, 2015.
[22] A. Saini, S. Daniel, S. Saini, and A. Mittal, "KannadaRes-NeXt: a deep residual network for Kannada numeral recognition," in *Studies in Big Data*, Springer Singapore, 2021, pp. 63–89, doi: 10.1007/978-981-15-9492-2_4.
[23] H. T. Chandrakala and G. Thippeswamy, "Deep convolutional neural networks for recognition of historical handwritten Kannada characters," in *Advances in Intelligent Systems and Computing*, vol. 1014, Springer Singapore, 2020, pp. 69–77, doi: 10.1007/978-981-13-9920-6_7.
[24] V. C. Hallur and R. S. Hegadi, "Handwritten Kannada numerals recognition using deep learning convolution neural network (DCNN) classifier," *CSI Transactions on ICT*, vol. 8, no. 3, pp. 295–309, Sep. 2020, doi: 10.1007/s40012-020-00273-9.
[25] S. T. Narasimhaiah and L. Rangarajan, "Deep convolutional neural networks for degraded printed kannada character recognition," *Indian Journal of Computer Science and Engineering*, vol. 12, no. 3, pp. 719–727, Jun. 2021, doi: 10.21817/indjcse/2021/v12i3/211203187.
[26] S. Vijayarani and A. Sakila, "Performance comparison of OCR tools," *International Journal of UbiComp*, vol. 6, no. 3, pp. 19–30, Jul. 2015, doi: 10.5121/iju.2015.6303.

BIOGRAPHIES OF AUTHORS

Sridevi Tumkur Narasimhaiah received her Master's degree in Computer Applications (MCA) in 2005 from Visvesvaraya Technological University, Belgaum. She worked as Assistant Professor in Department of MCA for 13 years in the field of teaching which includes a research experience of 6 years. She is currently, a research Scholar at Department of Studies in Computer Science, University of Mysore, ManasaGangothri Campus, Mysuru, working under the guidance of Dr. Lalitha Rangarajan. Her expertise is in C and C++ software developments. Her technical interests are programming and machine Automation. She can be contacted at email: tn.sridevi1@gmail.com.

Lalitha Rangarajan has been working as a Professor in the Department of Studies in Computer Science, University of Mysore, ManasaGangothri Campus, Mysuru. She has Master's degree in Mathematics from Madras University, India and from School of Industrial Engineering, Purdue, USA. Her career started with teaching mathematics and since 1988 shifted to Computer Science. She has received her Ph.D. degree in 2005 from the University of Mysore. She has over 36 years of teaching and 21 years of research experience and contributed to the field of Artificial Intelligence, Image Processing, Pattern Recognition, Cryptography, Computational Biology and Bioinformatics. She has served the University at various capacities. She has guided 13 Ph.D's and has about 100+ publications in peer reviewed journals and proceedings of conferences to her credit. She can be contacted at email: lali85arun@yahoo.co.in.