Degraded character recognition from old Kannada documents

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

Article Info

Article history:

Received Nov 19, 2020
Revised Jan 25, 2022
Accepted Feb 24, 2022

Keywords:

5-fold cross validation
Classifier validation methods
Optical character recognition
Degraded Kannada character recognition
Histogram of oriented gradients features
Preprocessed character database
Principal component analysis

ABSTRACT

This paper addresses preparation of a dataset of Kannada characters which are degraded and robust recognition of such characters. The proposed recognition algorithm extracts the histogram of oriented gradients (HOG) features of block sizes 4×4 and 8×8 followed by principal component analysis (PCA) feature reduction. Various classifiers are experimented with and fine K-nearest neighbor classifier performs best. The performance of proposed model is evaluated using 5-fold cross validation method and receiver operating characteristic curve. The dataset devised is of size 10,440 characters having 156 classes (distinct characters). These characters are from 75 pages of not well preserved old books. A comparison of proposed model with other features like Haar wavelet and Geometrical features suggests that proposed model is superior. It is observed that the PCA reduced features followed by fine K-nearest neighbor classifier resulted in the best accuracy with acceptance rate of 98.6% and 97.9% for block sizes of 4×4 and 8×8 respectively. The experimental results show that HOG feature extraction has a high recognition rate and the system is robust even with extensively degraded characters.

Corresponding Author:

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

1. INTRODUCTION

Optical character recognition (OCR) algorithms produce varying recognition accuracies for a variety of document types [1]. Some challenges in document image recognition are complex layouts, clear and degraded texts, characters with varying font sizes and styles, captured using different illumination levels/imaging devices and variety of degradation factors. These problems have led to the development of algorithms that suit a specific set of attributes in the document. In the proposed work, the investigations are carried out for achieving good recognition rates towards degraded characters inherent in old Kannada document images [2]. 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 suit the properties or characteristics of documents and a very few address the design needs of degraded old 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) [3]. 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 degradation of multiple character instances belonging to same class. Figure 1 shows sample of characters extracted Figure 1(a) before preprocessing and Figure 1(b) after preprocessing. It also shows degradation types covered for a specific class.

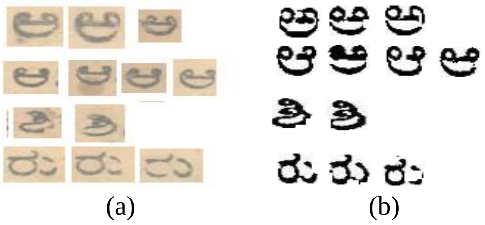


Figure 1. Sample of characters extracted (a) before preprocessing and (b) after preprocessing

In the literature, there are several attempts addressing variety of challenges for various types of documents during the recognition process of characters, some of the significant works are discussed here. Lakshmi [4] worked on degraded historical manuscripts for identification of Telugu characters from palm leaf characters by acquiring an additional 3D features from palm leaves. The proposed 3D depth sensing approach subtracts the background and extracts the text from degraded palm manuscripts. Feature extraction is achieved through histogram of oriented gradients (HOG) and cell-based directions followed by feature dimensionality reduction using differential evolution (DE) and standard particle swarm optimization (PSO) with best recognition rate as 93.1% using 'HOG' feature set. Sandhya and Krishnan [5] proposed a novel approach of rebuilding the broken Kannada characters and then used neural networks to classify the zonal features. The isolated characters are subject to the process of 50×50 normalization and character images are reduced to single pixel thickness. Reconstruction is performed by filling the gaps of the broken character using the end point algorithm. A recognition accuracy of 98.9% is achieved for broken characters on synthetically generated datasets. Experimentation on the real datasets segmented from historical document is not promising. Biswas *et al.* [6] presented a hybrid deep architecture for robust recognition of text lines of degraded printed documents with a moderately large annotated database created for degraded Bangla documents to study recognition. Gaussian mixture model based strategy is used for text extraction and a convolution neural network (CNN) with two layers of BLSTM cells and a connectionist temporal classification (CTC) layer have been used for recognition of the segmented text lines. The model provided 88.63% character level accuracy on the degraded image sample while existing Bangla OCR system provided 81% character level accuracy on the same image sample. Kumar *et al.* [7] proposed an efficient approach for automatic number plate recognition from low resolution images acquired using CCTV cameras. In this technique the input image is subject to an edge directed interpolation algorithm followed by a high pass Gaussian filter and maximally stable extremal regions (MSER) computation phases. Character recognition is performed through hybrid approach that uses template matching and achieved an average accuracy of 96.2%.

Trier *et al.* [8] surveyed various feature extraction methods for off-line recognition of segmented (isolated) characters. The paper discusses about feature extraction techniques and its applicability towards the different character representation forms. The paper also reveals properties of the extracted features for each feature extraction method and each character representation forms. Yang and Yan [9] have described a modified logical thresholding method for binarization of seriously degraded and very poor quality gray-scale document images. This method can deal with complex signal-dependent noise, variable background intensity caused by non-uniform illumination, shadow, smear or smudge and very low contrast. The run-length histogram for selected image regions and six local adaptive thresholding algorithms are employed for enhancement of various document images. Kunte and Samuel [10] devised a simple and efficient optical character recognition system for basic symbols (vowels and consonants) in printed Kannada text, which can handle different font sizes and font types. Features such as Hu's invariant moments and Zernike moments are classified using neural networks and achieved a good recognition rate of 96.8%.

Tonazzini *et al.* [11] proposed an integrated system for the processing of highly degraded printed characters that are de-noised through wavelet filtering, the text lines are detected, extracted, and segmented by adaptive thresholding and classified using feed forward multilayer neural network for character recognition. Experiments carried out showed the satisfactory performance. But the integration mechanism devised for its different blocks allows for obtaining a very precise segmentation of the single characters. Sagar *et al.* [12] devised an algorithm using brute force approach for character segmentation for Kannada OCR using database approach. Segmentation of lines and words are performed using horizontal and vertical projection profiles by dividing the character in to top, middle and bottom region. Support vector machine (SVM) classifier provided increased efficiency even when the character has a conjunct consonant. Recognition of degraded characters using dynamic Bayesian networks (DBNs) is investigated by Likforman-Sulem and Sigelle [13] using two HMM architectures coupled into a single DBN model which is an extension of one-dimensional hidden Markov models (HMM) to represent characters. The study showed coupled architectures work with better accuracy on the degraded characters than the basic HMM's. James *et al.* [14] have developed a Alexnet based Architecture for Malayalam handwritten character recognition using handcrafted feature extraction methods for classification. Multi class SVM is used for training the data and achieved better accuracy for 44 primary characters and 36 compound characters.

Kumar and Ramakrishnan [15] proposed an elegant method to recognize merged and split characters occurring in degraded printed documents. Their approach is a non-linear enhancement on histogram of the gray level degraded word images using power-law transformation (PLT) where fractional values of gamma are chosen automatically by their algorithm. So that the processed word images result in better connectivity with broken strokes during binarization. The technique has been applied on 1685 benchmark scanned datasets of degraded word images and noticed an absolute improvement of 14.8% in unicode level recognition accuracy of their SVM classifier. Ramteke *et al.* [16] have made an attempt to develop OCR framework for handwritten Marathi character recognition. New sine cosine algorithm is proposed for identifying the handwritten Marathi text using features like statistical, global transformation, geometrical and topological features. The weighted one-against-rest support vector machines (WOAR-SVM) is used for classification and achieved an accuracy of 95.14%.

Antony *et al.* [17] have devised a handwritten character recognition system on Tulu script. Haar features are extracted and the classification method is AdaBoost algorithm. Recognized characters are mapped to editable document of Kannada characters. This model seems to be the first attempt on handwritten script of Tulu. They have tested the model only with few samples. Thushara *et al.* [18] discuss feature extraction and classification. Features extracted are geometrical features and SVM classifiers employed for Malayalam handwritten character recognition. Geometrical features uses loops, endpoints and junction points to identify the Malayalam characters. In addition to this some more features such as orientations of arcs and different types of junction point features are also used. Diem and Sablatnig [19] focused on recognizing degraded handwritten characters using local features. Model is investigated on ancient manuscripts. The designed model is a segmentation free approach but based on local descriptors. Characters are recognized initially using SVM classifier and then by voting measure of neighboring local descriptors.

A novel character recognition approach has been proposed by Dutta *et al.* [20]. The approach results in decrease in error rate of 15% on highly degraded Indian language scripts. As OCR's performance is considerably good on good quality images, the model proposes character n-gram images, which are nothing but, grouping the consecutive character/component segments. Model successfully resolves the ambiguity between confusing characters in the recognition stage. The labels obtained from recognizing constituent n-grams are then combined to obtain a final label. When tested across English and Malayalam document images, model shows an improved recognition accuracy on highly degraded document images. Bar-Yosef *et al.* [21] propose a variation method based on segmentation of gray-scale images on highly degraded historical documents. Model accepts a training set of characters and constructs a small set of shape models that cover most of the training set's shape variance. Model accepts a gray-scale degraded image as input and constructs a custom made shape, which covers the shape models that best fits the character boundary. The model does not limit to any particular shape. Authors claim that proposed method yields highly accurate results in both segmentation and recognition on highly degraded character images. Average distance between the boundaries of respective segmented characters was 0.8 pixels (average size of the characters are 70×70 pixels).

Connected and degraded text recognition by Agazzi *et al.* [22] uses enhanced planar hidden Markov models (PHMMs). The approach automatically segments highly blurred and touching texts into characters as a part of recognition process. Experiments conducted on 24,000 highly degraded images of city names and a database of 6,000 images resulted in accuracy of 99.65% and 98.76%, whereas it is rejected by a high performance commercial OCR machine. Thillou *et al.* [23] developed an embedded application for degraded text recognition. Application has been built on mobile device, which gives access of text information to blind or visually impaired. The system needs 3 key technologies such as text detection, optical character recognition, and speech synthesis. Blind users and the mobile environment imply two strong constraints. First, pictures will be taken without control on camera settings and a priori information on text (font or size) and background. The second issue is to link several techniques together with an optimal compromise between computational constraints and recognition efficiency. Model presents the overall description of the system from text detection to OCR error correction.

Nomura *et al.* [24] presents a morphological approach to degraded character segmentation and feature extraction from highly degraded images. Algorithm depends on histogram which automatically detects fragments and merges these fragments before segmenting the fragmented characters. Morphological thickening automatically locates reference lines for dissecting the overlapped characters. Morphological thinning and segmentation cost calculation determines the base line for segmenting the connected characters. Model detects fragmented, overlapped, or connected characters and adaptively applies one of these algorithms without manual fine-tuning. Model considers some highly degraded images as license plate images for experimentation. The output of the model is a feature vector which keeps useful information to be used as input to an automatic pattern recognition system. Yokobayashi and Wakahara [25] follow a technique which applies a binarization and recognition on variety of degradations in complex backgrounds. Algorithm works in two stages. First one automatically selects one axis in RGB color space by a suitable threshold for segmentation. The other one is global affine transformation (GAT), which is nothing but distortion tolerant grayscale character recognition and this result in highest correlation value between input and template images. Model uses a 698 test images extracted from ICDAR 2003 datasets. Model achieves a performance of 81.4% ranging from 94.5% for clear images to 39.3% for seriously distorted images.

It is evident from literature that various attempts are reported on recognition of degraded characters with respect to various south Indian scripts. Most of the research outcomes report an accuracy rate of around 80-90%. Further, the techniques employed for enhancement of degraded characters are frequency based enhancement techniques in majority of works. Features like statistical, geometric, topological and hand-crafted features are highly employed for classification and recognition of characters. A system that is specifically designed for recognition of degraded printed Kannada characters is desired. The model specifically addressing the needs of handling variety of degradations particularly seen in ancient document images is not yet available. The proposed model is devised based on one of the efficient models (HOG features) that have shown the best performance in the literature. The novelty of this work lies in collecting several documents with variety of degraded characters. Documents are preprocessed and characters are extracted (as in OCR) using binary image analysis (BIA) [3].

The remaining part of the paper is organized as follows. Section 2 briefs the dataset creation. Section 3 describes the stages in the proposed approach, while section 4 is about the experimental analysis and an empirical proof of the efficiency of the proposed model compared to other existing feature extraction techniques. Finally, section 5 concludes the paper mentioning some drawbacks of the proposed method.

2. PREPARATION OF DATA

The detailed description on the datasets for degraded printed character recognition is explained in this section. Higher recognition accuracy is one of the key objectives of OCR system. In the proposed work, it is intended to realize higher accuracies for recognition of degraded Kannada character images extracted from ancient Kannada documents. The character images are extracted from the scanned image of ancient documents and subsequently subject to removal of variety of noises such as aging marks, bleed through, annotations and broken/dilated appearance of characters through morphological processing, followed by segmentation of connected components, and finally marginal noise removal of non-textual regions. The database of degraded characters is developed by considering characters of all possible degradations in each of the class. The character classes assumed for classification include only the frequently repeated characters in the degraded Kannada documents. The details of Kannada alphabetical set and the number of samples in each class is as follows. Kannada Varnamale consists of 49 letters of which 13 are called as swaras (vowels), 2 are Yogavaahakas (partly vowel and partly consonant) comprising of Anuswara- ಅಂ and Visarga- ಅಃ, followed by 25 are Vargeeya Vyanjanas (structured consonants) and 9 are Avargeeya Vyanjanas (unstructured consonants) as shown in Figure 2.

Characters from 75 pre-processed degraded pages from 10 books are extracted, out of which the frequently occurred characters from the set of vowels and consonants are considered for recognition. All the characters with frequency of occurrence as one and less than one is encircled in red ink in Figure 2. The frequency of occurrence of each of the characters listed in Figure 2 is shown in Figure 3(a) frequency of occurrence of vowels and Figure 3(b) frequency of occurrence of consonants. Along with the vowels and consonants, the extracted characters also include single level vowel consonant groups including all together the number of characters extracted from 75 documents crossed more than 16500 characters. The number of vowels, consonants and vowel consonant clusters or simple compound characters obtained includes 10,440 characters and remaining characters which are multi and complex compound are not included in the dataset. Few samples of severely degraded complex characters subsequent to preprocessing are shown in Figure 4. These characters include Figure 4(a) simple compound, Figure 4(b) multi compound and Figure 4(c) complex compound.

From the Figures 3(a) and 3(b), it is evident that there totally 40 classes of characters from frequently occurring vowels (9) and consonants (31). For each character, multiple variations of degradations are collected to create the reference samples for training so as to capture the intra class variations. Each degraded character sample in turn represented in 4 orientations (so that OCR can recognize documents in any direction) comprises into n×4 samples in each class, where n is the number of samples with variable degradations. Along with vowels and consonants there are vowel and consonant clusters which are called simple compound characters. The occurrence of the compound characters is less compared to vowels and consonants which are as shown in Table 1.

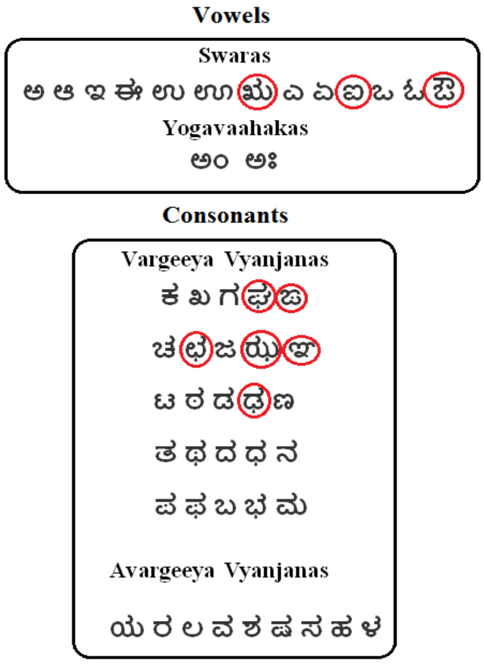
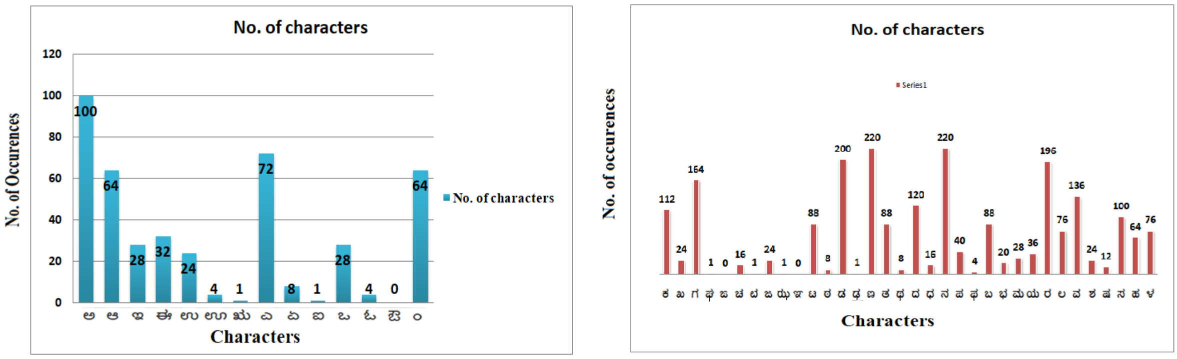
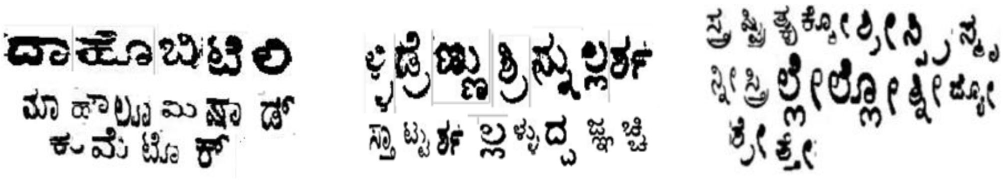


Figure 2. Kannada Varnamale



(a) (b)

Figure 3. Character-wise statistics (a) frequency of occurrence of vowels and (b) frequency of occurrence of consonants



(a) (b) (c)

Figure 4. Complex characters (a) simple compound, (b) multi compound, and (c) complex compound

Table 1. Frequency of occurrence of vowel and consonant clusters

Class Label	Character	Character Occurrences	Class Label	Character	Character Occurrences	Class Label	Character	Character Occurrences
41	PÁ	80	80	y	4	119	AiÀiï	16
42	Q	120	81	xÉ	16	120	AiÀiÁ	36
43	PÀÄ	20	82	zÁ	48	121	¬Ä	60
44	PÀÆ	24	83	¢	136	122	AiÀÄÄ	120
45	PÉ	8	84	zÀÄ	140	123	AiÀÄ	12
46	PÉÆ	12	85	zÀÆ	24	124	AiÉÆ	28
47	PË	28	86	zÉÆ	12	125	gï	32
48	T	100	87	zË	116	126	gÁ	40
49	SÉ	28	88	zsÁ	112	127	K	88
50	UÁ	100	89	¢ü	12	128	gÀÄ	52
51	V	4	90	zsË	4	129	gÀÆ	20
52	UÀÄ	140	91	£ï	144	130	gÉ	20
53	UÀÆ	28	92	£Á	20	131	gÉÆ	24
54	UÉ	20	93	¤	116	132	gË	120
55	UË	88	94	£ÀÄ	136	133	¯ï	120
56	WÁ	160	95	£ÀÆ	124	134	¯Á	88
57	X	120	96	£É	24	135	°	140
58	WÉ	88	97	¥Á	8	136	®Ä	136
59	WÉÆ	24	98	¦	120	137	®Æ	140
60	ZÁ	128	99	¥ÀÄ	44	138	¯É	16
61	a	120	100	¥É	116	139	¯ÉÆ	4
62	ZÀÄ	88	101	¥ÉÆ	112	140	ªÁ	124
63	ZÀÆ	84	102	¥sÁ	4	141	«	16
64	ZÉ	140	103	¦ü	8	142	ªÀÅ	116
65	eÁ	40	104	¥sÉ	4	143	ªÀÇ	112
66	f	44	105	¨Á	212	144	±Á	132
67	dÄ	40	106	©	128	145	²	16
68	eÉ	112	107	§Ä	20	146	±ÀÄ	36
69	eÉÆ	16	108	§Æ	152	147	µÁ	148
70	mÁ	60	109	¨É	20	148	¹	128
71	n	120	110	¨ÉÆ	196	149	¸ÀÄ	124
72	lÄ	80	111	¨sÁ	20	150	¸É	8
73	r	112	112	©ü	8	151	¸ÉÆ	24
74	qÀÄ	28	113	¨sÀÄ	74	152	ºÁ	68
75	vÁ	40	114	¨sÀÆ	16	153	»	120
76	w	68	115	ªÀiï	8	154	ºÀÆ	16
77	vÀÄ	20	116	ªÀiÁ	20	155	½	68
78	vÉ	4	117	«Ä	4	156	¼É	32
79	vÉÆ	92	118	ªÀÄÆ	88			

3. PROPOSED METHOD

The proposed model for classification of degraded Kannada characters begins with acquisition of characters extracted from training datasets, followed by feature extraction and dimensionality reduction. The processing stages of the proposed approach for degraded character recognition from old Kannada documents is as shown in Figure 5. 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 K-nearest neighbor (WKNN), cosine KNN, cubic KNN, and ensemble Adaboost classifiers. The detailed process of feature extraction using HOG is discussed here.

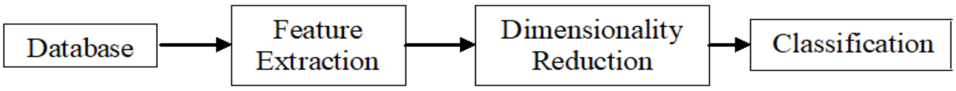


Figure 5. Block diagram for classification of degraded Kannada characters

3.1. Feature extraction using histogram of oriented gradients (HOG) and dimensionality reduction

HOG is widely used pattern descriptor for various pattern recognition problems. These features capture shape information precisely. As the Kannada characters are shape sensitive, HOG features are used in the proposed model. The HOG descriptors are highly advantageous due to individual feature description in local cells, and being invariant to geometric and photometric transformations, except for object orientation. 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. 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. Considering 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. The computed gradient angles are distributed into 9 bins (0-20, 20-40, …, 160-180). Making blocks of images reduces the number of features significantly.

HOG computes one or more orientations of character partition in each cell. The orientations computed from each block (cell) are recorded in the respective bins. Thus, a block of size 4×4 is described by 9 features. There are 49 such non-overlapping 4×4 blocks. 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 36×36 (1296) features describing a character. Figure 6 describes the process of extracting the HOG features and classification protocol.

As degradations increase the complexity of recognition resulting in higher error rates, it is required to build the database for training by including characters with different amount of degradations. HOG features are further reduced using principal component analysis (PCA). Upon using PCA data will be represented in a better form with minimal loss of information, as the goal of PCA is to reduce the dimensionality of the data while retaining as much information as possible in the original dataset. PCA also helps to build more effective data analysis on the reduced dimensional space. The resultant feature vectors are submitted to the classifier for recognition of character class.

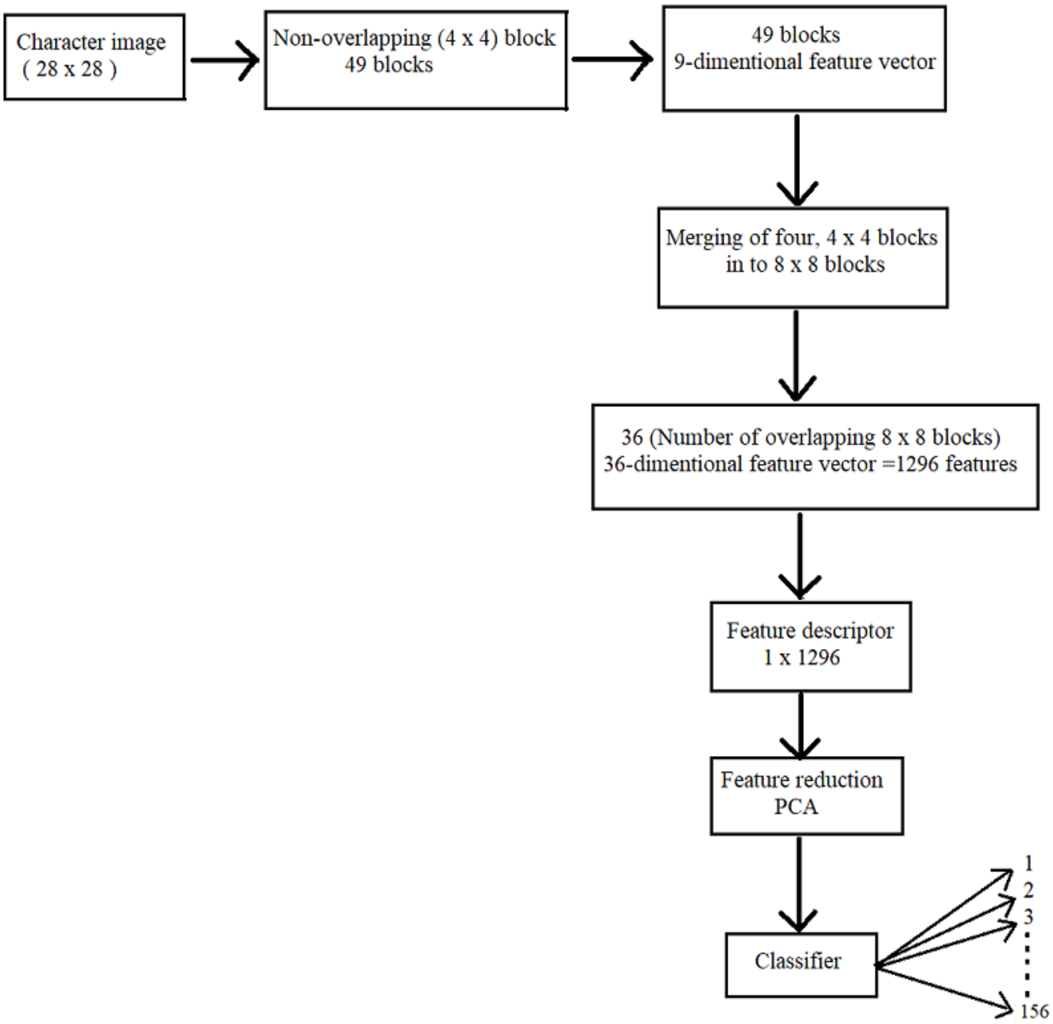

Character image
(28 x 28)
Non-overlapping (4 x 4) block
49 blocks
49 blocks
9-dimentional feature vector
Merging of four, 4 x 4 blocks
in to 8 x 8 blocks
36 (Number of overlapping 8 x 8 blocks)
36-dimentional feature vector =1296 features
Feature descriptor
1 x 1296
Feature reduction
PCA
Classifier
1
2
3
156

Figure 6. HOG feature extraction and classification

4. EXPERIMENTAL ANALYSIS

As discussed in the dataset preparation section, the training and testing is conducted on the overall datasets of about 10,440 character instances covering 156 classes with 5-fold cross validation technique. Each class consists of varying number of characters of multiple degradations. The experimentation is conducted using machine learning to determine the robust classification technique through comparative study. HOG features extracted for each cell sizes (4×4 and 8×8) are evaluated separately with different classifiers using 5-fold cross validation method.

Features extracted using HoG feature descriptor are subjected to dimensionality reduction using principal component analysis (PCA) as the feature set size is too large. Experiments on reduction using PCA with various variance levels suggest that PCs covering 85% variation yields optimal classification results. For cell size 4×4, 1296 features are being reduced to 89 features covering 85% of total variance and whereas 144 features have been reduced to 19for cell size of 8×8. The outcome of PCA with the most prominent features needed for object recognition are trained separately for cell sizes 4×4 and 8×8 using fine Gaussian support vector machines, fine KNN, medium KNN, WKNN, cosine KNN, cubic KNN, and ensemble Adaboost. Table 2 shows the performance metrics of classification accuracy.

Table 2. Classification accuracy of classifiers vs. cell sizes with HOG features

Cell size	Fine Gaussian SVM	Fine KNN (k=1)	Medium KNN (k=10)	WKN (k=10)	Cosine KNN (k=10)	Cubic KNN (k=10)	Ensemble Adaboost Classifier
4×4	98.3%	98.6%	91.4%	98.4%	91.6%	91.2%	31.1%
8×8	97.3%	97.9%	89.7%	97.2%	89.2%	89.4%	11.3%

It may be observed that a remarkable accuracy of 98.6% and 97.9% is achieved with fine KNN classifier with k=1 for cell sizes 4×4 and 8×8. Observe that Adaboost performance is very poor. This may be because many of the individual classifiers perform poor. SVM failed to attain convergence between the test subsets and training subsets of data, as the features need to be scaled. Due to very large scale variance, SVM failed to provide appreciable results. As fine KNN resulted in best performance with reduced feature set, the performance of this KNN is compared with the classifiers with other features.

Haar wavelet features (Antony *et al.* [17]) and geometrical features (Thushara *et al.* [18]) are computed and the same set of classifiers is used. Table 3 represents the classification accuracy of Haar wavelet and geometrical features in the dataset of degraded Kannada characters. These features are used as it is and not transformed or reduced.

Table 3. Classification accuracy of classifiers with Haar wavelet and geometric features

Features type	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
Haar Wavelet feature	69.2%	73.1%	56.8%	72.3%	57.4%	56.8%	12.9%
Geometrical feature	94.8%	97.4%	81.7%	96.6%	80.1%	81.6%	19.1%

It is noticed clearly from the Table 3, there is decline in accuracy of all classifiers with the use of these features and also it indicates that performance is better with fine KNN with both HOG and these features. The performance of a classification model is often measured through receiver operating characteristic curve (ROC). Here the true positive rate (TPR) is plotted against false positive rate (FPR). The area under the curve (AUC) is calculated. AUC represents the probability that a random positive sample is positioned to the right of a random negative sample which ranges from 0 to 1. If the learned model’s predictions are 100% wrong, it will have an AUC of 0.0 where as the one whose predictions are 100% correct will have an AUC of 1.0. Table 4 shows the number of classes with AUC >=85%. Table 5 shows the number of classes with AUC =1. Observe that Geometrical feature appear to be best performer when AUC >=85% is the performance measurement. However, if the number of classes with AUC 1 is the criteria, HOG outperforms appreciably as compared to all other methods including Geometrical features. Table 6 shows the accuracy of online OCR and fine KNN.

From Table 6 it is observed that the there is a steep drop in the accuracy of online OCR when degraded characters are tested across 30 samples. Most of the degraded characters are miss-recognized as different characters and ASCII symbols by the current available online OCR [26]. Proposed HOG tested across 10,440 samples yielded far more accuracy rate as compared to OCR.

Table 4. Classes with AUC >=85%

Features type		>=85% TPR out of 156 classes
HoG	4×4	142
	8×8	143
Haar Wavelet feature		73
Geometrical feature		153

Table 5. Classes with AUC =1

Features type		100% TPR out of 156 classes
HoG	4×4	114
	8×8	103
Haar Wavelet feature		73
Geometrical feature		153

Table 6. Accuracy of online OCR

		Accuracy of Fine KNN (k=1)	Accuracy of online OCR with 30 samples
HoG	4×4	98.6%	3.3%
	8×8	97.9%	

5. CONCLUSION

The proposed work addresses the preparation of data of degraded Kannada characters as well as the use of HOG features for the recognition of degraded characters. A new dataset encompassing variety of degradations and multiple instances of characters belonging to same class is devised. Further, the datasets are subject to experimentation using HOG feature descriptors followed by dimensionality reduction. From the results of classification experimented using fine Gaussian support vector machines, fine K-nearest neighbor (KNN), medium KNN, WKNN, cosine KNN, cubic KNN, and ensemble Adaboost classifiers, it is noticed that Fine KNN is the best performer compared to other classifiers with high acceptance rate. A limitation of the proposed work is that multi and complex compound characters are not part of the database as these require a lot of additional work in the segmentation stage. This will be our future work.

REFERENCES

[1] M. Meshesha and C. V Jawahar, “Optical character recognition of amharic documents,” *African Journal of Information & Communication Technology*, vol. 3, no. 2, Aug. 2007, doi: 10.5130/ajict.v3i2.543.

[2] B. V. Dhandra, P. Nagabhushan, M. Hangarge, R. Hegadi, and V. S. Malemath, “Script identification based on morphological reconstruction in document images,” in *18th International Conference on Pattern Recognition (ICPR’06)*, 2006, pp. 950–953, doi: 10.1109/ICPR.2006.1030.

[3] 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 (IJRTE)*, vol. 8, no. 4, pp. 6660–6669, 2019, doi: 10.35940/ijrteD9101.118419.

[4] T. R. Vijaya Lakshmi, “Reduction of features to identify characters from degraded historical manuscripts,” *Alexandria Engineering Journal*, vol. 57, no. 4, pp. 2393–2399, Dec. 2018, doi: 10.1016/j.aej.2017.09.009.

[5] N. Sandhya and R. Krishnan, “Broken kannada character recognition — a neural network based approach,” in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, Mar. 2016, pp. 2047–2050, doi: 10.1109/ICEEOT.2016.7755047.

[6] C. Biswas, P. S. Mukherjee, K. Ghosh, U. Bhattacharya, and S. K. Parui, “A hybrid deep architecture for robust recognition of text lines of degraded printed documents,” in *2018 24th International Conference on Pattern Recognition (ICPR)*, Aug. 2018, pp. 3174–3179, doi: 10.1109/ICPR.2018.8545409.

[7] T. Kumar, S. Gupta, and D. S. Kushwaha, “An efficient approach for automatic number plate recognition for low resolution images,” in *Proceedings of the Fifth International Conference on Network, Communication and Computing - ICNCC ’16*, 2016, pp. 53–57, doi: 10.1145/3033288.3033332.

[8] Ø. Due Trier, A. K. Jain, and T. Taxt, “Feature extraction methods for character recognition-a survey,” *Pattern Recognition*, vol. 29, no. 4, pp. 641–662, Apr. 1996, doi: 10.1016/0031-3203(95)00118-2.

[9] Y. Yang and H. Yan, “An adaptive logical method for binarization of degraded document images,” *Pattern Recognition*, vol. 33, no. 5, pp. 787–807, May 2000, doi: 10.1016/S0031-3203(99)00094-1.

[10] R. Sanjeev Kunte and R. D. Sudhaker Samuel, “A simple and efficient optical character recognition system for basic symbols in printed Kannada text,” *Sadhana*, vol. 32, no. 5, pp. 521–533, Oct. 2007, doi: 10.1007/s12046-007-0039-1.

[11] A. Tonazzini, S. Vezzosi, and L. Bedini, “Analysis and recognition of highly degraded printed characters,” *International Journal on Document Analysis and Recognition*, vol. 6, no. 4, pp. 236–247, Apr. 2003, doi: 10.1007/s10032-003-0115-y.

[12] B. M. Sagar, G. Shobha, and P. R. Kumar, “Character segmentation algorithms for Kannada optical character recognition,” in *2008 International Conference on Wavelet Analysis and Pattern Recognition*, Aug. 2008, pp. 339–342, doi: 10.1109/ICWAPR.2008.4635800.

[13] L. Likforman-Sulem and M. Sigelle, “Recognition of degraded characters using dynamic Bayesian networks,” *Pattern Recognition*, vol. 41, no. 10, pp. 3092–3103, Oct. 2008, doi: 10.1016/j.patcog.2008.03.022.

[14] A. James, M. J, and C. Saravanan, “Malayalam handwritten character recognition using AlexNet based architecture,” *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, vol. 6, no. 4, pp. 393–400, Dec. 2018, doi: 10.52549/ijeei.v6i4.518.

[15] H. R. S. Kumar and A. G. Ramakrishnan, “Gamma enhanced binarization - an adaptive nonlinear enhancement of degraded word images for improved recognition of split characters,” in *2019 National Conference on Communications (NCC)*, Feb. 2019, pp. 1–6, doi: 10.1109/NCC.2019.8732254.

[16] S. P. Ramteke, A. A. Gurjar, and D. S. Deshmukh, “A novel weighted SVM classifier based on SCA for handwritten marathi character recognition,” *IETE Journal of Research*, pp. 1–13, Jun. 2019, doi: 10.1080/03772063.2019.1623093.

[17] P. J. Antony, C. K. Savitha, and U. J. Ujwal, “Haar features based handwritten character recognition system for Tulu script,” in *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, May 2016, pp. 65–68, doi: 10.1109/RTEICT.2016.7807784.

[18] K. Thushara, A. James, and C. Saravanan, “Feature extraction using geometrical features for Malayalam handwritten character recognition system,” in *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, Mar. 2017, pp. 477–482, doi: 10.1109/WiSPNET.2017.8299802.

[19] M. Diem and R. Sablatnig, “Recognition of degraded handwritten characters using local features,” in *2009 10th International Conference on Document Analysis and Recognition*, 2009, pp. 221–225, doi: 10.1109/ICDAR.2009.158.

[20] S. Dutta, N. Sankaran, K. P. Sankar, and C. V. Jawahar, “Robust recognition of degraded documents using Ccharacter N-grams,” in *2012 10th IAPR International Workshop on Document Analysis Systems*, Mar. 2012, pp. 130–134, doi: 10.1109/DAS.2012.76.

[21] I. Bar-Yosef, A. Mokeichev, K. Kedem, I. Dinstein, and U. Ehrlich, “Adaptive shape prior for recognition and variational segmentation of degraded historical characters,” *Pattern Recognition*, vol. 42, no. 12, pp. 3348–3354, Dec. 2009, doi: 10.1016/j.patcog.2008.10.005.

[22] O. E. Agazzi, S.-S. Kuo, E. Levin, and R. Pieraccini, “Connected and degraded text recognition using planar hidden Markov models,” in *IEEE International Conference on Acoustics Speech and Signal Processing*, 1993, pp. 113–116, doi: 10.1109/ICASSP.1993.319760.

[23] C. Thillou, S. Ferreira, and B. Gosselin, “An embedded application for degraded text recognition,” *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 13, pp. 1–9, Dec. 2005, doi: 10.1155/ASP.2005.2127.

[24] S. Nomura, K. Yamanaka, O. Katai, H. Kawakami, and T. Shiose, “A novel adaptive morphological approach for degraded character image segmentation,” *Pattern Recognition*, vol. 38, no. 11, pp. 1961–1975, Nov. 2005, doi: 10.1016/j.patcog.2005.01.026.

[25] M. Yokobayashi and T. Wakahara, “Binarization and recognition of degraded characters using a maximum separability axis in color space and GAT correlation,” in *18th International Conference on Pattern Recognition (ICPR’06)*, 2006, pp. 885–888, doi: 10.1109/ICPR.2006.326.

[26] V. S and S. A, “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, Manasa Gangothri 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, Manasa Gangothri 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 PhD 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 12 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.