Malaya Journal of Matematik, Vol. 9, No. 1, 751-756, 2021

https://doi.org/10.26637/MJM0901/0132

Review of offline handwritten text recognition in south Indian languages

A. T. Anju[1*], Binu P. Chacko[2] and K. P. Mohammad Basheer[3]

Abstract

Offline Handwritten character recognition is a popular and challenging area of research under pattern recognition and image processing. In this article, offline handwriting recognition methods performed in south Indian languages including Telugu, Tamil, Kannada and Malayalam are presented. A description about south Indian languages and an overview of general handwriting recognition systems are also presented briefly. Convolutional Neural Networks (CNNs) and classifier combination methods have provided better performance among proposals provided by the researchers.

Keywords

CNN, deep learning, Offline Handwriting Recognition, South Indian Script.

[1,3] *Department of Computer Science, Sullamussalam Science College, Kerala-673639, India.*

[2] *Department of Computer Science, Prajyoti Niketan College, Kerala-680301, India.*

*Corresponding author: [1] anjuat89@gmail.com

Article History: Received **19** December **2020**; Accepted **02** March **2021**

©2021 MJM.

Contents

1. Introduction

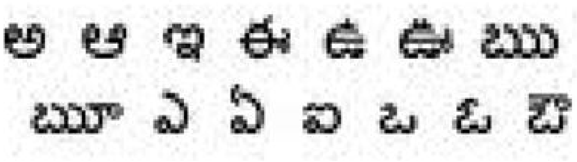
Handwriting recognition (HWR) is a challenging component in Optical Character Recognition (OCR) and it is also called intelligent character recognition (ICR). HWR is a fast growing area of research due to its high impact in business areas and reducing the precious man time. It has applications in the areas like banking, healthcare, insurance, education etc. In the area of document image processing, HWR is a vital component [1]. Handwriting recognition can be categorized into two, online handwriting recognition and offline handwriting recognition. Online handwriting recognition is performed while we are interacting with the electronic device; so the temporary information like coordinate pair of the touching of the electronic pen and the pen trajectory movements are available in the case of online handwriting recognition process. Offline handwriting recognition is performed on the scanned image which contains handwritten text to recognize, that is, the handwritten characters in a scanned image are converted to digital form and can be stored in the computer for future use. Offline character recognition is more complex and challenging not only because of the cursive and complex structure of handwriting but the unavailability of temporal information also [2].

Offline handwriting recognition is the process of converting the text in a digitized image into a letter code format which can be processed by the computer. Mainly two approaches are used for the recognition process such as analytical approach and holistic approach. In an analytical approach the whole word is divided into different characters and then the feature extraction is performed. In the case of holistic approach the features are extracted from the whole word [3]. The different steps involved in the process of offline handwritten character recognition are preprocessing, feature extraction, segmentation, classification and recognition. The applications in this area include postal automation, automatic form processing, historical document recognition, automatic bank cheque processing, etc. [4].

The primary objective of this article is to present an all-inclusive review of the state of the art in offline handwriting recognition systems for South Indian scripts. For this purpose, we studied the research papers on offline handwritten text recognition in south Indian languages, and the present paper is organized in 8 sections based on the information gathered. In section 2, we have presented the characteristics of south Indian languages. In section 3, the overview of the architecture of handwriting recognition systems is discussed. Sections 4, 5, 6, and 7 deal with the study of handwriting text recognition in Telugu, Tamil, Kannada, and Malayalam languages. Section 8 concludes the paper.

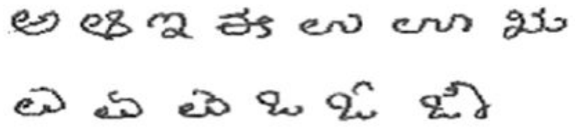
2. Characteristics of south indian languages

In South India, the people speak Dravidian languages. The four major Dravidian languages in south india are Telugu, Tamil, Kannada, and Malayalam. Telugu language uses Telugu script for writing and spoken mostly in the states of Andhra Pradesh, Telangana and Yanam District of the Union Territory of Pondicherry. Telugu language is spoken by 82 million people across the world. Telugu script consists of 35 consonants and 18 vowels. The figure shows the Telugu vowels.



In South India, Tamil is the most widely used spoken language and is the official language in the states of Tamil Nadu and the union territory of Puducherry. Tamil is spoken by the people in Telangana, Kerala, Andhra Pradesh, Karnataka, and Andaman and Nicobar Islands. The number of speakers of Tamil language is 77 million people across the world. Tamil language uses Tamil script for writing. Tamil script consists of 12 vowels and 23 consonants. The figure shows the Tamil vowels.

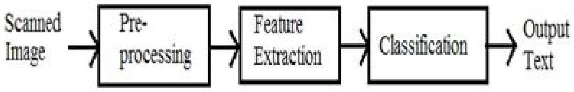
Kannada language, primarily written in Kannada script, and is spoken by 44 million people across the world. It is derived from Sanskrit, and spoken in the state of Karnataka and border areas of Andhra Pradesh, Tamil Nadu and Maharashtra. There are 16 vowels consonants in Kannada script. The figure shows the Kannada vowels.



Malayalam language is written in Malayalam script, which is a Brahmic script. It is the principal language in the state of Kerala, union territories of Lakshadweep and Puducherry. Malayalam language is spoken by 38 million people across the world. Malayalam script consists of 13 vowels and 36 consonants. The figure shows the Malayalam vowels.

3. General steps in handwriting recognition system

The major steps in the handwriting recognition process are preprocessing, feature extraction and classification. Some recognition systems avoid the segmentation step and directly the features are extracted.



3.1 Preprocessing

Preprocessing is a very important step in image processing performed to get an enhanced image or to extract some useful information from the image for a better recognition system. Different preprocessing steps performed on the scanned image are document image binarization, normalization, noise reduction, skew and slant correction, segmentation, etc.

3.2 Feature Extraction

Feature extraction is the process of extracting the features from the input image for the classification process. Feature extraction is the most important step for achieving high recognition accuracy. In deep learning, the model performs the feature extraction automatically.

3.3 Classification

In classification, the input image is assigned to a proper class which it belongs to based on the features extracted from the input image. Different classification techniques are Logistic regression, Naive Bayes classification, Support Vector Machine (SVM), K-nearest neighbour (K-NN), etc. Recently, deep learning has been used for classification for better performance.

4. Handwriting recognition studies in telugu language

Ganji et al. proposed a deep learning model for handwriting recognition by creating a transfer learning model named "TCR_new_model.h5" from the VGGNET model. The dataset contained 1400 unique characters and for the experimentation the dataset was divided into two, training and testing using VGG-16 and classification was performed using VGGNET-16 and achieved 92% accuracy [5]. Dara et al. described a method for handwriting recognition using SVM classifier. In the preprocessing stage, binarization was performed using thresholding and also performed rotation to augment the dataset. In this, the features were extracted using 2-dimensional Fast Fourier Transform (2D-FFT). For experimentation 3,750 samples were used for training and 500 samples for testing and achieved 71% accuracy [6]. Sastry et al. proposed a method based on zoning features. Binarization and normalization were performed in preprocessing for better feature extraction. Then the image was converted to 100 zones and the features were extracted from the zones by calculating the sum of all pixel intensities in the zone. Then converted the 100 features in to column matrix and performed the same for the training and testing dataset. The recognition was performed by comparing the Euclidean distance between column matrices. The dataset contained 18,750 samples for training, and 500 samples for testing and achieved 78% accuracy [7]. Manisha et al. proposed a glyph segmentation method. The different preprocessing steps performed on the input image were binarization, skew and slant correction, character segmentation and thinning. Various approaches were performed for identifying the connected and unconnected glyphs to segment into top vowel ligature glyphs, main glyphs, bottom vowel ligature glyphs and consonant conjuncts glyphs. The experiment was performed on 2,576 characters samples and the system achieved 91.14% success rate [8]. Vijay and Rajeshwara proposed a method for handwriting recognition using Multi-Layer Perceptron (MLP). In preprocessing, binarization, skeletonization and normalization were performed on the input image. The MLP classifiers contained a two-layer feed forward network with nonlinear sigmoidal functions and the MLP classifiers in the hidden layer were trained using error back propagation algorithms. The experiments were performed on 195 samples in which 95 samples were used for training and 100 samples for testing, and achieved 84.9% accuracy [9].

5. Handwriting recognition studies in tamil language

Kavitha and Srimathi proposed a method for offline handwritten character recognition using CNN. The main two parts in the model were training and testing, and in training, the model was trained using the preprocessed data. In recognition, first the preprocessing was performed and then the recognition performed using the trained model. The proposed CNN model contained 9 layers including convolutional layers, max pooling layers and fully connected layers. The experiments were performed on HP Labs character dataset containing 82,929 images and achieved 97.7% accuracy with dropout [10]. Pragathi et al. described a method for Tamil character recognition using deep learning. VGG 16 CNN was used in this study which consisted of 13 convolution layers with pooling layers in between them, then Loss 3 classifier layer and output layer. The experiments were performed on a dataset containing 15,600 images and achieved 94.52% accuracy [11]. Kowsalya and Periasamy proposed a neural network model for handwritten Tamil character recognition with a modification using elephant herding optimization algorithm. In preprocessing, noise reduction was performed using Gaussian filter, and then binarization using thresholding method and skew detection were performed. In segmentation, paragraph, line and word segmentations were performed using spatial space detection technique, vertical histogram technique and horizontal histogram technique respectively. Then features were extracted on individual image glyph and then the features were fed to the neural network for recognition. The weights in the neural network were optimized using Elephant Herding Optimization technique to improve the recognition accuracy. In this, the dataset contained 150 images and achieved 93% accuracy [12].

Deepa and Rao proposed a method based on high dimensional features using nearest interest point (NIP) classifier. In this, the feature extraction was performed using speeded up robust features (SURF) to generate interest points (IP) from the image's physical structure. Then the features were fed to the nearest interest point (NIP) classifier for recognition. The experiments were performed on HP Labs India's hpl-tamil-iso-char dataset, they selected 156 symbols for the database creation, and 500 samples were created for each symbol. The system achieved 90.2% accuracy [13]. Prakash and Preethi described a method using deep CNN architecture. The CNN contains two convolutional layers followed by max pooling layers, the two fully connected layers and an output layer with Rectified Linear Unit (ReLU) activation function. In this, Nesterov Accelerated batch gradient descent optimization algorithm was used to improve accuracy and dropout regularization technique was used to avoid over fitting. For testing and training experiments, the IWFHR database, containing 124 classes was used and achieved the accuracy of 88.2% and 71.1% respectively [14]. Jayakanthan et al. proposed a method for handwriting recognition using an Artificial Neural Network (ANN) called Residual Neural Network (ResNet). In preprocessing, noise reduction, skewing, smoothing and resizing were performed on the input image. The preprocessed data were fed to the Resnet50 model to train the model and the data after preprocessing were fed to the trained model for recognition. For the experiment, this study derived a dataset from Hp Labs India's HPL-Tamil-Iso-Char database and made it available in UNIPEN format of 15,000 images, and the model achieved 96% accuracy [15]. Vinotheni et al.

described a method for handwriting recognition using Modified Convolutional Neural Network (M-CNN). The M-CNN architecture contained a convolution layer, pooling layer and fully connected layer in a single sequence. They created a dataset containing 156 distinct characters with 350 samples for each class and achieved 97.07% of accuracy [16].

6. Handwriting recognition studies in kannada language

Ramesh et al. proposed a method for handwritten word and character recognition using SVM with CNN. In preprocessing, binarization was performed using the Otsu method, noise elimination by median filter and segmentation using the methods called vertical projection profile and the bounding box. CNN contained convolutional and subsampling sub layers which performed feature extraction, and the activation function ReLU was used for triggering. The tenth level of CNN contained L2-SVM for classification. The experiments were performed on several printed and handwritten databases and achieved a recognition accuracy of above 80% [17]. Rao et al. described a method for handwriting recognition using CNN. In preprocessing, grayscale conversion was performed to reduce the complexity of the image and noise reduction was performed using Non-local means algorithm to reduce Gaussian noise. Then contrast normalization was performed to enhance the quality of the image and binarization was performed using thresholding method. Segmentation was performed using OpenCV contour function used to recognize lines, words and characters and also performed the augmentation to increase the diversity of the image. Feature extraction was performed using CNN and the features were fed to the Artificial Neural Network (ANN) to perform the classification. The experiments were performed on Chars74K dataset and achieved 95.11% for the training set and 86% for the testing set [18].

Ramesh et al. presented a method for handwriting recognition using CNN. In preprocessing, resizing and grayscale conversion were performed. Feature extraction and classification were performed by CNN. In this, the dataset was divided into two, for training and testing, containing 13566 images in training set and 3399 images in testing set of the same size and achieved an accuracy of 78.73% [19]. Karthik and Murthy proposed a method for the recognition of handwritten characters using Deep Belief Network (DBN).

Distributed average of gradients (DAG) features was extracted from the image for classification and the recognition was performed by DBN. The proposed DBN architecture contains five hidden layers and one classification layer. The restricted Boltzmann machine (RBM) present in the DBN was trained using Contrastive Divergence (CD) algorithm. The experiments were performed on a dataset containing 18,800 samples and achieved 97.04% accuracy [20]. Joe et al. proposed a method for handwriting character recognition using Convolution Neural Network (CNN). In preprocessing, the grayscale image from the dataset was converted to an intermediary file then resizing and normalization were performed. The feature extraction was performed by CNN and the extracted features were fed to a dense Artificial Neural Network (ANN) for classification. For experimentation, the Char74k dataset consisting 657 Kannada characters with of 25 samples for each were used, and the system achieved 57% accuracy [21]. Belagali and Angadi proposed a method for handwriting recognition using a Probabilistic Neural Network (PNN) classifier. In preprocessing, the background removal was performed using thresholding and then character segmentation was performed using connected components labeling. The segmented characters were subdivided into three horizontal zones, and Hu's invariant moments, Horizontal and Vertical profile features were extracted from these three zones. Then a pattern dictionary was created for the classification and the recognition was performed using PNN with 94.69% accuracy [22].

Angadi and Sharanabasavaraj described a method for handwritten Kannada character recognition using SVM. Image resizing and thinning operations were performed in the preprocessing stage. In feature extraction, structural features were extracted and the feature vector was fed to the SVM classifier for classification. For the experiment, 49 symbols from the Kannada script were selected and 50 samples were collected for each symbol. The system achieved 87.49% accuracy [23]. Bannigidad and Gudada proposed a method for the recognition of historical Kannada handwritten document images based on Histogram of Oriented Gradients (HOG) feature descriptors. The dataset was created by collecting 1200 images of historical handwritten Kannada documents of different age types namely Hoysala, Vijayanagara and Mysore dynasties. In preprocessing, binarization and block-wise segmentation were performed on the input image. K-nearest neighbour (K-NN) and SVM classifiers were used for the classification. K-NN achieved 92.3% accuracy and SVM achieved 96.7% accuracy [24].

7. Handwriting recognition studies in malayalam language

Chacko and Dhanya described a model for the recognition of offline handwritten Malayalam character using multiple classifiers. At first the preprocessing was performed, and it contained binarization using Otsu's method; segmentation was performed using horizontal and vertical projection method and the input image was normalized using bicubic interpolation. In feature extraction phase two types of features, gradient features and density features were extracted and for the classification the gradient features were fed to a neural network and the density features were fed to another neural network and the two feed forward neural networks were combined using the combination schemes Max rule, Sum rule, Product rule and Borda count method. A dataset consisting of 825 samples written by 25 individuals was used for the experiment and they achieved 81.82% accuracy using the product rule combination scheme [25]. Varghese et al. proposed a three stage feature extraction technique for recognizing handwritten character in Malayalam language. In preprocessing phase, binarization, segmentation and thinning were performed on the input image. Then, the feature extraction was performed in three stages. In the first stage the geometrical features like corner, ending, bifurcation and loop were considered and similar shaped characters were grouped into a class and in the second stage, feature extraction techniques for recognizing each character within that class was performed. In the third stage, rules regarding the formation of Malayalam word were considered for checking the probability of occurrence of the character in that position based on moment variant features, and the system achieved better recognition accuracy [26].

Nair et al. proposed a method for Malayalam handwritten character recognition using CNN. At first, preprocessing was performed to enhance the quality of the input image for better recognition accuracy. For the database creation, the first 6 characters from the Malayalam alphabet were used and then the data augmentation was performed and 2 lakh samples were created. In this, feature extraction was also performed by CNN and there were no handcrafted features. Then the CNN modeling was performed to fix the number of layers in convolutional layers, max pooling layers, ReLU layers and fully connected layers [27]. Jino et al. proposed a method for handwriting recognition using deep CNN with transfer learning strategy. In this, CNN was used to perform feature extraction and SVM for classification. The experiments were performed on a Malayalam dataset containing 10,676 handwritten samples and achieved 96.90% accuracy [28].

8. Conclusion

In this paper, offline handwriting recognition systems for handwritten Telugu, Tamil, Kannada and Malayalam are presented. Various preprocessing and feature extraction techniques are carried out by the researchers, and in the case of deep learning methods, diverse features are automatically extracted by the model. More studies are carried out using CNNs and from the study we conclude that the CNNs performed better for the recognition process.

References

[1] Bhowmik, S., Malakar, S., Sarkar, R., Basu, S., Kundu, M. and Nasipuri, M. Off-line Bangla handwritten word recognition: a holistic approach. *Neural Computing and Applications*, 31(10)(2019), 5783-5798.

[2] Tiwari, U., Jain, M. and Mehfuz, S. Handwritten character recognition—an analysis. *In Advances in System Optimization and Control,* (pp. 207-212). Springer, Singapore. (2019).

[3] Ghosh, M., Malakar, S., Bhowmik, S., Sarkar, R. and Nasipuri, M. Feature selection for handwritten word recognition using memetic algorithm. *In Advances in Intelligent Computing*, (pp. 103-124). Springer, Singapore. (2019).

[4] Roy, P. P., Bhunia, A. K., Das, A., Dey, P. and Pal, U. HMM-based Indic handwritten word recognition using zone segmentation. *Pattern Recognition*, 60(2016), 1057-1075.

[5] Ganji, T., Velpuru, M. S., and Dugyala, R. Multi Variant Handwritten Telugu Character Recognition Using Transfer Learning. In IOP Conference Series: *Materials Science and Engineering,* 1042(1)(2021), (012026). IOP Publishing.

[6] Dara, R. and Panduga, U. Telugu handwritten isolated characters recognition using two dimensional fast fourier transform and support vector machine. *International Journal of Computer Applications,* (0975 – 8887), 116(5)(2015).

[7] Sastry, P. N., Lakshmi, T. V., Rao, N. K., Rajinikanth, T. V. and Wahab, A. *Telugu handwritten character recognition using zoning features.* In 2014 International Conference on IT Convergence and Security (ICITCS) (pp. 1-4). IEEE. (2014).

[8] Manisha, C. N., Krishna, Y. S. and Reddy, E. S. Glyph segmentation for offline handwritten Telugu characters. *In Data Engineering and Intelligent Computing,* (pp. 227-235). Springer, Singapore. (2018).

[9] Rao, K. V. K. R. R. Improvement in efficiency of recognition of handwritten Telugu script. *International Journal of Inventive Engineering and Sciences* (IJIES), 2(1)(2013), 1-4.

[10] Kavitha, B. R. and Srimathi, C. Benchmarking on offline handwritten Tamil character recognition using convolutional neural networks. *Journal of King Saud University-Computer and Information Sciences*, (2019), (1319-1578).

[11] Pragathi, M. A., Priyadarshini, K., Saveetha, S., Banu, A. S. and Aarif, K. M. Handwritten Tamil character recognition using Deep Learning. *In 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN),* (2019) (1-5). IEEE.

[12] Kowsalya, S. and Periasamy, P. S. Recognition of Tamil handwritten character using modified neural network with aid of elephant herding optimization. *Multimedia Tools and Applications*, 78(17)(2019), 25043-25061.

[13] Deepa, R. A. and Rao, R. R. A novel nearest interest point classifier for offline Tamil handwritten character recognition. *Pattern Analysis and Applications*, 23(1)(2020), 199-212.

[14] Prakash, A. A. and Preethi, S. Isolated offline Tamil handwritten character recognition using deep convolutional neural network. *In 2018 International Conference on Intelligent Computing and Communication for Smart World*, (I2C2SW), (2018) (278-281). IEEE.

[15] R. Jayakanthan, A. H. Kumar, N. Sankarram, B. S. Charulatha, and A. Ramesh. Handwritten Tamil Character Recognition Using ResNet. *In International Journal*

of Research in Engineering, Science and Management, 3(3)(2020), 133-137.

[16] Vinotheni, C., Pandian, S. L. and Lakshmi, G. Modified convolutional neural network of Tamil character recognition. *In Advances in Distributed Computing and Machine Learning,* (pp. 469-480). Springer, Singapore. (2020).

[17] Ramesh, G. and Kumar, S. *Recognition of Kannada handwritten words using SVM Classifier with Convolutional Neural Network.* In 2020 IEEE Region 10 Symposium (TENSYMP) (pp. 1114-1117). IEEE. (2020).

[18] Rao, A. S., Sandhya, S., Anusha, K., Arpitha, C. N. and Meghana, S. N. Exploring Deep Learning Techniques for Kannada Handwritten Character Recognition: A Boon for Digitization. *International Journal of Advanced Science and Technology,* 29(5)(2020), 11078-11093.

[19] Ramesh, G., Sharma, G. N., Balaji, J. M. and Champa, H. N. *Offline Kannada handwritten character recognition using Convolutional Neural Networks*. In 2019 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE) (pp. 1-5). IEEE. (2019).

[20] Karthik, S. and Murthy, K. S. Deep belief network based approach to recognize handwritten Kannada characters using distributed average of gradients. *Cluster Computing*, 22(2)(2019), 4673-4681.

[21] Joe, K. G., Savit, M. and Chandrasekaran, K. *Offline character recognition on segmented handwritten Kannada characters*. In 2019 Global Conference for Advancement in Technology (GCAT) (pp. 1-5). IEEE. (2019).

[22] Belagali, N. and Angadi, S. A. OCR for handwritten Kannada language script. *Int. J. Recent Trends Eng. Res.(IJRTER)*, 2(08)(2016), 190-197.

[23] Angadi, S. A. and Angadi, S. H. Structural features for recognition of hand written Kannada characters based on SVM. *International Journal of Computer Science, Engineering and Information Technology*, 5(2)(2015), 25-32.

[24] Bannigidad, P. and Gudada, C. Age-type identification and recognition of historical kannada handwritten document images using HOG feature descriptors. *In Computing, Communication and Signal Processing*, (pp. 1001-1010). Springer, Singapore, (2019).

[25] Chacko, A. M. M. and Dhanya, P. M. Multiple classifier system for offline Malayalam character recognition. *Procedia Computer Science*, 46(2015), 86-92.

[26] Varghese, K. S., James, A. and Chandran, S. A novel tri-stage recognition scheme for handwritten Malayalam character recognition. *Procedia Technology*, 24(2016), 1333-1340.

[27] Nair, P. P., James, A. and Saravanan, C. Malayalam handwritten character recognition using convolutional neural network. 2017 *International Conference on Inventive Communication and Computational Technologies* (ICICCT), 278–281. (2017). DOI: 10.1109/ICICCT.2017.7975203.

[28] Jino, P. J., Balakrishnan, K. and Bhattacharya, U. Offline handwritten Malayalam word recognition using a deep architecture. *In Soft Computing for Problem Solving*, (pp. 913-925). Springer, Singapore. (2019).

⋆⋆⋆⋆⋆⋆⋆⋆⋆
ISSN(P):2319 – 3786
Malaya Journal of Matematik
ISSN(O):2321 – 5666
⋆⋆⋆⋆⋆⋆⋆⋆⋆