RESEARCH ARTICLE

DOI: 10.34256/ijcci2122

International Journal of Computer Communication and Informatics

Offline Recognition of Malayalam and Kannada Handwritten Documents Using Deep Learning

Ayna Asokan [a], Sreeleja N Unnithan [a,*]

[a] Department of Electronics and Communication Engineering, NSS College of Engineering, Palakkad, Kerala, India.

* Corresponding Author: sreeleja@rediffmail.com

Received: 03-08-2021, Revised: 01-10-2021, Accepted: 05-10-2021, Published: 30-10-2021

Abstract: For a variety of reasons, handwritten text can be digitalized. It is used in a variety of government entities, including banks, post offices, and archaeological departments. Handwriting recognition, on the other hand, is a difficult task as everyone has a different writing style. There are essentially two methods for handwritten recognition: a holistic and an analytic approach. The previous methods of handwriting recognition are time- consuming. However, as deep neural networks have progressed, the approach has become more straightforward than previous methods. Furthermore, the bulk of existing solutions are limited to a single language. To recognise multilanguage handwritten manuscripts offline, this work employs an analytic approach. It describes how to convert Malayalam and Kannada handwritten manuscripts into editable text. Lines are separated from the input document first. After that, word segmentation is performed. Finally, each word is broken down into individual characters. An artificial neural network is utilised for feature extraction and classification. After that, the result is converted to a word document.

Keywords: Holistic, Analytic, Handwriting recognition

Introduction

The digital revolution has ushered in a technology-driven economy in which digital documents provide various advantages in terms of information preservation over traditional media. It has features such as security, ease of access, and keyword-based search. Instead of putting a document on paper, it can be safely kept and conveniently accessed when it is digitalized. Such transfers aid many government agencies that deal with a lot of handwritten material. Post offices, banks, and other financial institutions are examples of entities that process huge amounts of historical data in the form of manuscripts. Other applications include automatic number plate recognition, document security, and so on. As a result, handwriting recognition has become an area of study, and the government has backed widespread document digitization. As

a result, recognition of handwriting is more useful.

The purpose of a handwriting recognition system is to convert human readable characters into a machine editable version. It includes two types: online and offline. When a user writes on a screen of an electronic device that recognises handwriting, the information is immediately gathered and stored in an online character recognition system. Characters can be changed in real time in the online mode. Data is scanned after it is written on paper in an offline system. In online word recognition, the direction and arrangement of strokes are key aspects. It is more difficult to recognise offline word recognition because it relies on photos scanned from handwritten documents. Handwritten text is much more difficult to recognise than printed writing. The difference between characters in printed text is small. Handwritten texts, on the other hand, are not the same. Handwriting comes in a variety of styles. A lot of factors influence a person's writing style, including the writer's emotion, mood, age, and health status, as well as the variation afforded by the pen employed. All of these characteristics influence the way you write.

Offline handwritten word recognition can be divided into two categories: Analytic and Holistic. In the analytic technique, words are divided into characters or sub characters, and each character is identified one by one. A number of word recognition systems have been built using this strategy. One of the most difficult components of this approach is determining correct segmentation points from the handwritten cursive words. A holistic approach, on the other hand, eliminates the need for character segmentation. It considers the full sentence. The characteristics are extracted directly from the word images in this approach, and then classification methods are used to recognise it. Even if the writing is poor, a comprehensive approach may have beneficial results. It does, however, have limitations, such as a predefined vocabulary size. This thesis takes an analytical approach. In most circumstances, literature considers only one language. The languages considered in this thesis are Malayalam and Kannada.

Malayalam is the official language of Kerala and one of India's 22 scheduled languages. This language is based on the Grantha script, which is a descendant of Ancient Brahmi. The character set consists of 51 letters, with 13 vowels and 37 consonants. A total of 12 vowel marks are included in the set. The usage of these characters in diverse forms distinguishes the old script, on the other hand. The original Malayalam character set is highly complex. Because printing Malayalam was difficult, in the 1970s and 1980s, a reduced or reformed version of the script was introduced.

Karnataka's official language is Kannada. Kannda is based on the Bramhi script. Kannada is written with 52 letters, 16 vowels, and 36 consonants. The scripts also comprise ten alternative Kannada numerals for the decimal number system. Modifiers for consonants and vowels are also distinct symbols that alter the base sounds. The numbers for these modifications are the same as for the base characters. The symbols for consonants, consonant modifiers (optional), and vowel modifiers are graphically combined according to a set of principles to generate these characters, known as aksharas. There are a lot of characters in both languages. Also, there are a lot of compound characters in the character set. There is also a higher level of character likeness. All of these factors can add to the difficulty of character recognition.

Customized features can be added to the Artificial Neural Network which used for feature extraction and classification, to improve recognition accuracy.

Literature Review

The use of elliptical features and MLP-based classifiers as part of a holistic technique to recognise handwritten city names is proposed in [1]. A set of 65 elliptical properties is extracted from the word image. A hypothetical ellipse has been added to the word image. 13 global feature values are extracted from each word image. Because each handwritten word has a different shape and pixel distribution in differ- ent sub regions, computing its elliptical characteristics can successfully distinguish one handwritten word from another. A suitable classifier was picked based on their recognition accuracy on the test data set. They also used a three-fold cross validation approach for the recognition task. Despite the fact that word images are accurately recognised, skewness causes certain word samples to be wrongly classified. It was also revealed that the pixel distributions of words from different classes are almost identical.

[2] employs the tetragonal feature, a shape-based feature descriptor. Additionally, an elliptical feature and a vertical pixel density histogram-based feature are used. It aids in the capture of the contour or geometric quality of a handwrit- ten-word image. A fivefold and threefold cross-validation technique using MLP and SVM classifiers are employed to solve the recognition problem. The system displayed better recognition accuracy with the exception of a few word classes. Differences in spelling, complex shapes, and words with similar shapes from other classes resulted in misclassifications. In order to extract the features, [3] retrieves gradient orientation information from each of the word graphics. At the start, each word image is separated into an equal number of grids. Gradient-based properties have been retrieved from each of these grids. The Histogram of Oriented Gradients (HOG) feature descriptor is utilised as a local feature extractor. The classifier is Sequential Minimal Optimization (SMO). Despite having the best recognition accuracy, 5-fold cross validation had a number of problems. The classifier took longer to train since the feature vector was so huge. To get around this, certain feature dimension reduction techniques are to be employed. Low-level data from the word image is recovered in [4]

Density, pixel ratio, area, centroid, aspect ratio, projection length, and longest run are the derived features. The characteristics are obtained from a word image or sub-images. Each word graphic yields a feature vector with 89 elements.

The Arnold transform is used to confuse the word image in [5] Directional features are extracted in this system. The Hough transform is used to generate directional properties. The stroke orientation distribution of the cursive word determines the directional qualities. The Arnold transform is used first to estimate stroke orientation, followed by the Hough transform. For identifying Arabic handwritten text, [6] suggested a successful multiple classifier technique. Chebyshev moments with statistical and contour-based characteristics are employed for word recognition. The data was classified using a number of classifiers, including Support Vector Machine, Multi-Layer Perceptron, and Extreme Learning Machine.

The H-WordNet model, which uses a neural network, is proposed by [7] This model consists of five learnable layers, four convolutional layers, and one fully connected layer. The primary advantage of using a neural network is that it eliminates the need for humans to extract features. Dimensionality reduction and classification can also be eliminated. This model learns features on its own. This allows to save a substantial amount of time. It is implemented using the back-propagation algorithm. It creates a reliable system for recognising handwritten words. Stochastic gradient descent with momentum optimizer was used to train the parameters.

In [8] statistical characteristics are extracted. There are two Indian scripts under consideration. The handwritten text's skew is first discovered and corrected. After preprocessing, a headline estimate is performed. The words are then separated into suitable pseudo-characters. The following step is to distinguish three statistical traits. They also used convolutional neural network-based transfer learning architectures and compared them to classic ones. The pre-trained designs included AlexNet, Resnet18, Resnet50, VGG-16, Google net, Densenet201, and VGG-19. This method was used to handle the handwritten words' slant. Despite its inconsistency, it is able to recognised the headline.

In [9] an effective character segmentation strategy for the Hindi language was proposed. The cursive writing of the script is also taken into account. The structural patterns are used to segment the data. There are three major steps to the process. In the first phase, the header line is removed. The header line is extracted in the initial step. The top strip is separated from the rest of the piece. It results in vertically divided middle and bottom zone components. It could be shadow characters, touching characters, characters with a lower modifier, conjuncts, or a mixture of these. In the second phase, the upper modifier is segmented by accumulating statistical data on intermediate individual components. In the third phase, these statistical data are used to choose components that require further segmentation. This approach can deal with skewed header lines and different writing styles.

[10] proposes a strategy for segmenting words and characters in Devanagari scripts. The methods used are Pixel Plot and Trace, as well as Re-plot and Retrace (PPTRPRT). It extracts the text section of the document. Along with skew and de- skew activities, iterative approaches for line segmentation are used. In pixel-space based word segmentation, these iteration findings are used. After word segmentation, characters are separated from words. In a variety of methods, the PPTRPRT approach can be used to distinguish characters from hand- written script. It also performs a number of normalization algorithms when writing to account for pen width and slant deviations.

Wavelet properties are extracted in [11] Feature extraction and classification are the two stages of this method. To extract features, the Haar wavelet transform is utilised. A Support Vector Machine classifier is used for classification. From the input image, a grayscale image is produced. Normalization of sizes is required. There are four sub images at LL1, LH1, HL1, and HH1. The next level decomposition is used to create the image at LL2, LH2, HL2, and HH2. Decomposition is carried out up to the third level. The retrieved feature vector is subsequently trained and tested using the Support Vector Machine classifier.

Two distinct classifiers are used in [12] The ensemble technique is applied when numerous classifiers are used to solve a problem. It has the ability to improve the performance of the system. All of the SURF (Speeded-UP Robust Features), curvature, and diagonal features are retrieved. As classifiers, support vector machines and neural networks are used.

Methodology

Image acquisition, pre-processing, segmentation, neural net- work training and testing, and finally conversion to word format are the six phases of the proposed system. The block diagram of the suggested handwriting recognition system is shown in figure1.

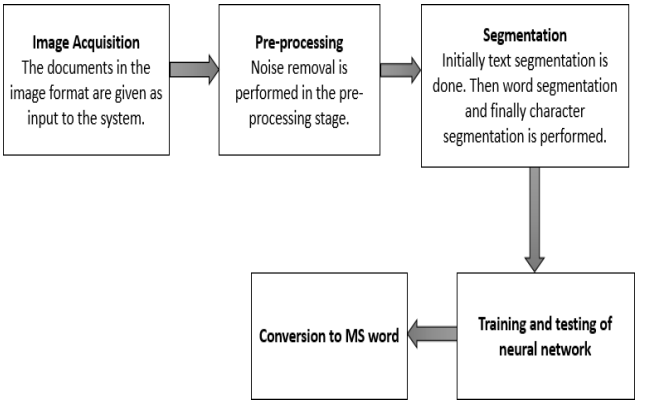

Figure 1. Block diagram of the proposed handwriting recognition system

A. Image acquisition

The document's image is included in the input. The image can be saved as a jpeg or a png file. A scanner is used to create the images.

B. Pre-processing

To reduce aberrations in the scanned image, preprocessing is used. Distortions might arise as a result of a low-quality scanner or document degradation. The scanned image is initially turned to grayscale. After that, Otsu's global thresholding approach is used to transform the grayscale image to binary. It's the most popular binarization method. Otsu's thresholding approach works by iterating over all potential threshold values and calculating a measure of spread for the pixel levels each side of the threshold, i.e. the pixels that either lie in foreground or background. The conversion with this method does not require any prior knowledge of the image. The goal is to achieve the smallest total threshold value for foreground and background spreads. The goal of the method is to reduce within-class variance while increasing between-class variance. The black pixels in the binarized image have pixel value zero, while the white pixels have pixel value one. These values are inverted for further processing. Finally, preliminary masking is applied to remove noise from the paper, such as unwanted dots.

C. Segmentation

Individual characters are extracted from the handwritten image using segmentation. The first step is to segment the lines. After that, the words are segmented. Finally, character segmentation is carried out, and the features of letters are examined in order to classify it.

1) *Line segmentation:* Line segmentation is first conducted on the preprocessed document. The Horizontal Projection [HP] method [13] is used. Document is divided into many rows. The sum of each row's pixel values are calculated. The Horizontal Projection method is named by the fact that the value is calculated row by row. The line boundaries for segmentation must be identified. The line borders are the rows in which the total pixel value is zero. The top and bottom of a line will have pixel values of zero. The first row with a non-zero HP value, followed by rows with a zero HP value, is then considered as the first segmentation point. The row with HP value zero, which comes after these non-zero HP values, is then picked as the final segmentation point. As a result, the initial and end points serve as segmentation line borders.
2) *Word segmentation:* To segment words from lines, the Vertical Projection [VP] method is employed [13] The number of black pixels in each column is determined once each line is divided into columns. The principle behind word segmentation is that the distance between words is greater than the distance between characters. The VP approach is applied and then the count of consecutive zeroes are calculated. Finally, the calculated value is compared to the predetermined threshold. If it is found to be more than the threshold value, it is taken a word boundary.
3) *Character segmentation:* The VP method is used for character segmentation. However, Connected Component Analysis [CCA] is also employed in conjunction with the VP approach [5] In a binary image, CCA returns all of the characters. This strategy can improve accuracy while also maintaining the character's order. It analyses an image for pixels with comparable pixel intensity values and groups them togethe. Following the grouping, each pixel is assigned a grey level or colour. The VP value is obtained first. The number of columns with VP value zero is discovered. Then, after the column with zero VP value, the first succeeding column is picked as the final segmentation point. If there isn't a vertical barrier between the letter and the connected component, CCA is used [14-19].

D. Training and testing

A database has been developed for training and testing. In an A4 page, 100-200 instances of each letter in Malayalam and Kannada are written. For training and testing, the database is scanned, and images of each letter are obtained. For training and testing, an Artificial Neural Network [ANN] is used. There are just minor changes between several letters in the Malay- alam and Kannada languages. This can result in classification errors. As a result, several custom features are retrieved, such as area, width and height ratios, and so on. This feature set is provided to the network to make the learning process of the network easier.

E. Conversion to Ms word

Finally, the document's image is converted to Word format. Some Ms Word library files are supplied for this. For the first two languages, a font type is chosen. The fonts chosen for Malayalam and Kannada are Baraha Kan New and Manorama, respectively. After that, character mapping is performed to verify the character's alternate code [keyboard input] in the installed font type.

Result and Discussion

A. Simulation Result

MATLAB Version 18 is used to run the simulation. The performance plot is shown in figure 2, which is plotted betweeen mean square error and epoch.

The regression plot is shown in figure 3. It is the graph that shows the expected output vs the desired value.

A scanned image of a Malayalam document is shown in figure

4. Figure 5 shows the output after masking.A line divided from the document is presented in figure 6. The words are then split from the line, as shown in figure 7. Figure 8 shows the result of character segmentation. The transformed document in word format is shown in figure 9.

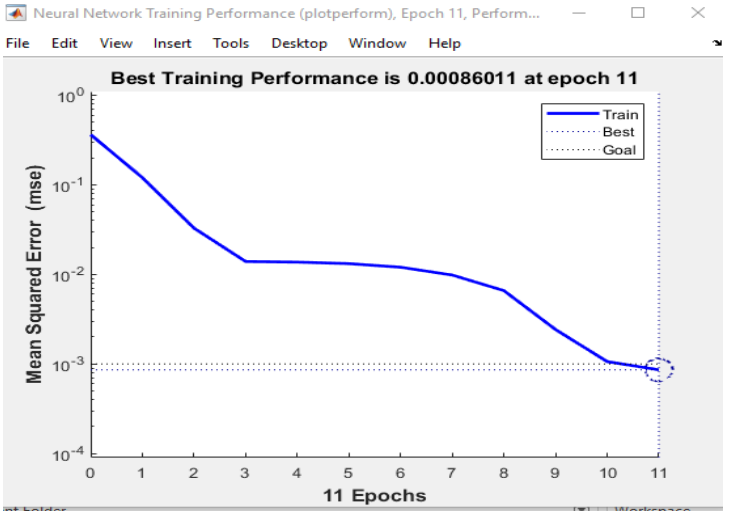

Figure 2. Performance plot

Figure 10 depicts a sample kannada document. Figure 11 depicts the masked output. A segmented line from the document is shown in figure12, while its word segmentation is shown in figure13.

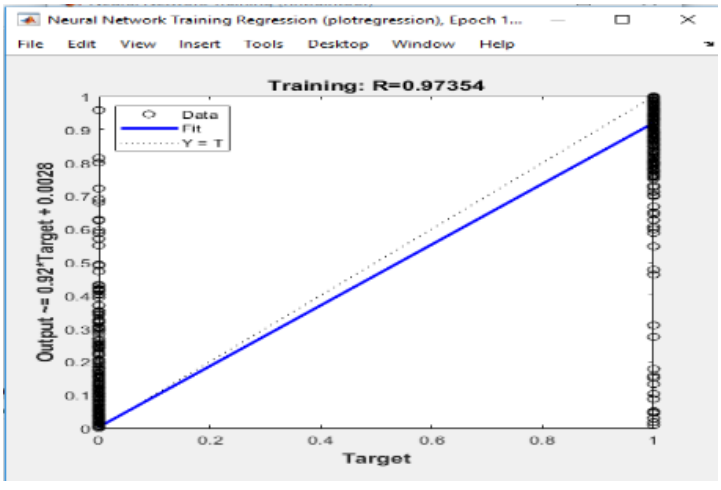

Figure 3. Regression plot

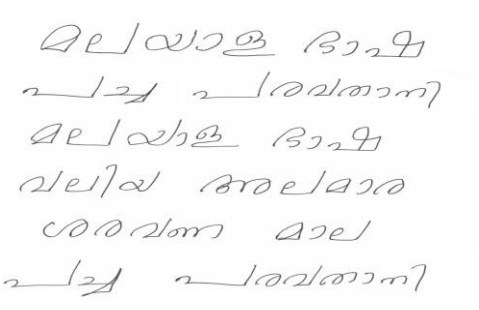

Figure4. Malayalam sample doc2

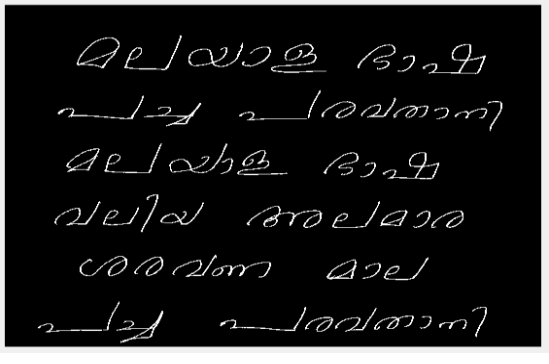

Figure 5. Output after masking

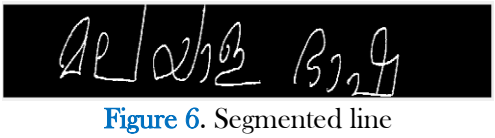

Figure 6. Segmented line

Figure 7. Word segmentation

In figure. 14, the character segmentation of the words presented in figure. 13 is shown. Figure 15 depicts the modified word document.

The malayalam text with several compound letters are shown in figure 16. A few letters from an old script are also displayed. The document's disguised output is shown in figure 17. Finally, the letters are recognised, and the associated word document is shown in figure 18.

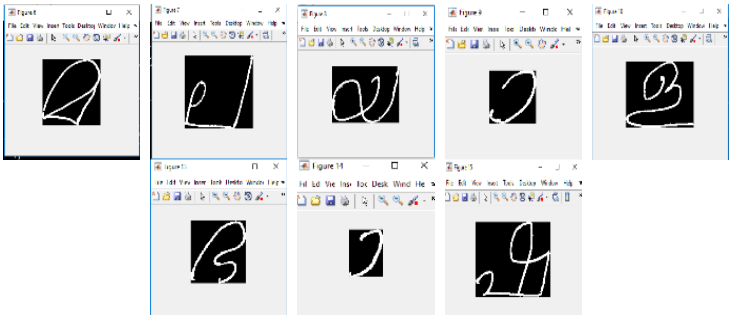

Figure 8. Character segmentation

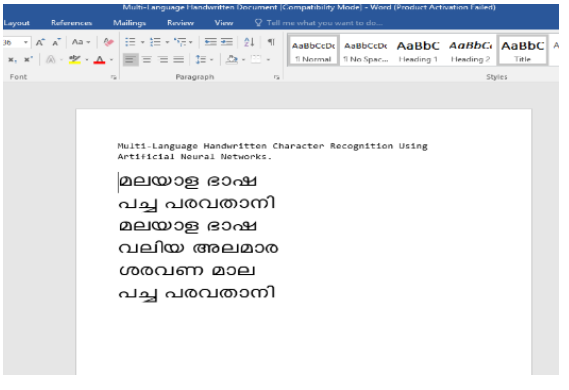

Figure 9. Malayalam Word doc2

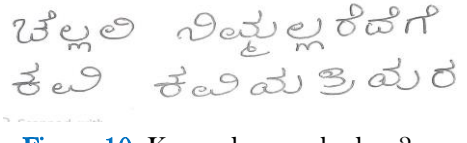

Figure 10. Kannada sample doc 3

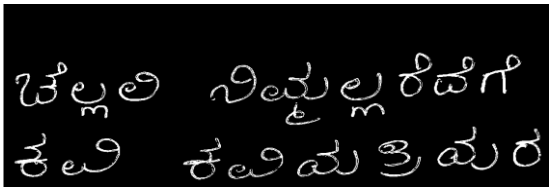

Figure 11. Output after masking

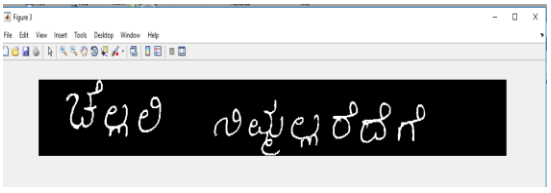

Figure 12. Segmented line

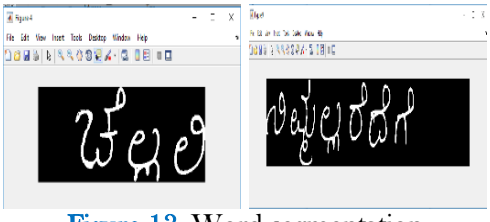

Figure 13. Word segmentation

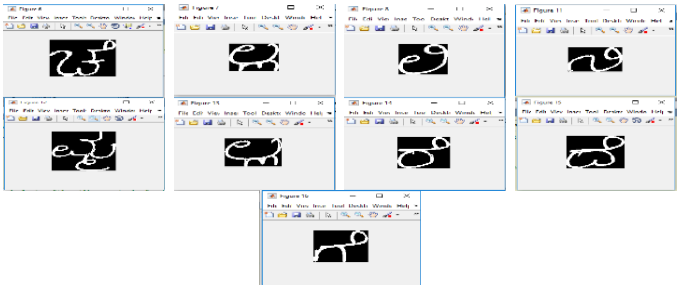

Figure 14. Character segmentation

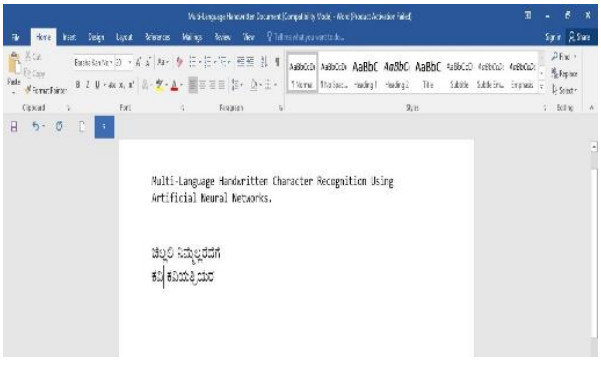

Figure 15. Kannada word doc3

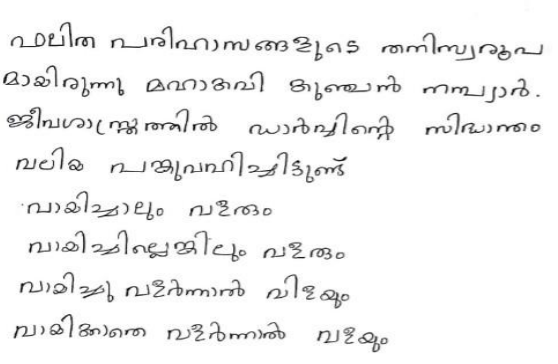

Figure 16. Malayalam doc4

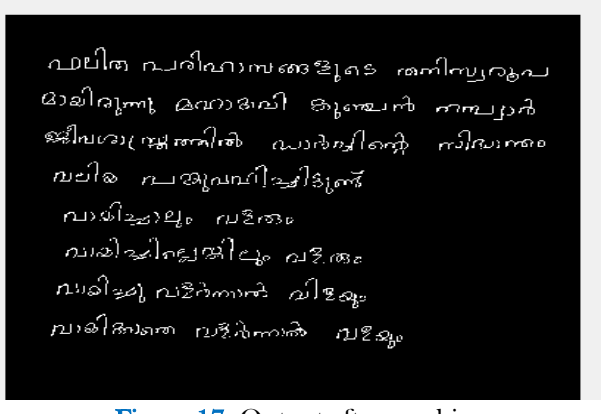

Figure 17. Output after masking

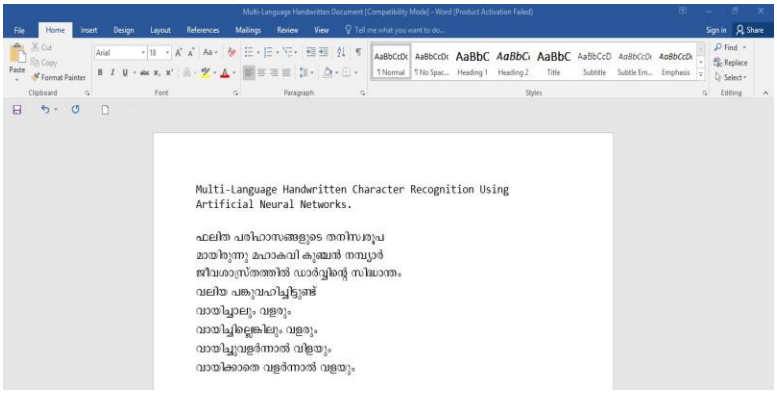

Figure 18. Word output

Conclusion

In this paper, an artificial neural network-based technique for recognising handwritten documents in Malayalam and Kannada is proposed. The network has been trained and tested successfully. The recognition accuracy was determined to be 96%. Various documents with

different handwritings were successfully detected and converted to word format. Researchers are still attempting to design a system that offers the best results because handwritten script recognition is a challenging and emerging topic. There are only a few works of literature that employ many languages. As a result, a real-time handwritten script recognition system capable of recognising a wide range of languages is needed.

References

[1] Bhowmik T.K., Parui S.K., Roy U., (2008). Discrim- inative hmm training with ga for handwritten word recognition, *2008 19th International Conference on Pattern Recognition,* IEEE, USA. https://doi.org/10.1109/ICPR.2008.4761830

[2] Rahiman M.A., & Rajasree M., (2011). Recognition of simple and conjunct handwritten malayalam characters using lcpa algorithm, *International Conference on Advances in Computing and Communications*, Springer, Berlin. https://doi.org/10.1007/978-3-642-22720-2_31

[3] Hashrin C., Jossy A., Sudhakaran K., Thushara A., & John A., (2019). Segmenting characters from malayalam hand- written documents, *In 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)*, IEEE, India.

[4] John J., Pramod K.V., & Balakrishnan K., (2012). Uncon- strained handwritten malayalam character recognition using wavelet transform and support vector machine classifier, *Procedia Engineering*, 30, 598–605. https://doi.org/10.1016/j.proeng.2012.01.904

[5] Dhaka V.P., Sharma M.K., (2015). An efficient segmenta- tion technique for devanagari offline handwritten scripts using the feedforward neural network, *Neural Computing and Applications*, 26(8), 1881–1893. https://doi.org/10.1007/s00521-015-1844-9

[6] Malakar S., Sharma P., Singh P.K., Das M., Sarkar R., & Nasipuri M., (2017). A holistic approach for handwritten hindi word recognition, *International Journal of Com- puter Vision and Image Processing (IJCVIP)*, 7(1), 59–78. https://doi.org/10.4018/IJCVIP.2017010104

[7] Saha S., Som T., (2011). Hand written character recognition using fuzzy membership function, *IJETSE International Journal of Emerging Technologies in Sciences and Engineering*, 5(2), 11-15. https://dx.doi.org/10.2139/ssrn.2009131.

[8] Sahoo S., Nandi S.K., Barua S., Bhowmik S., Malakar S., Sarkar R., (2018). Handwritten bangla word recognition using negative refraction-based shape trans- formation, *Journal of Intelligent & Fuzzy Systems*, 35(2), 1765–1777. https://doi.org/10.3233/JIFS-169712

[9] Cherkauer B.S., & Friedman E.G., (1995). A unified design methodology for cmos tapered buffers, *IEEE Transactions on Very Large-Scale Integration (VLSI) Systems*, 3(1), 99–111. https://doi.org/10.1109/92.365457

[10] Das D., Nayak D. R., Dash R., Majhi B., & Y.-D. Zhang, (2020). H-wordnet: A holistic convolutional neural network approach for handwritten word recognition, *IET Image Processing*, 49(9), 1794-1805. https://doi.org/10.1049/iet-ipr.2019.1398

[11] Bhowmik S., Malakar S., Sarkar R., & Nasipuri M., (2014). Handwritten bangla word recognition using elliptical features, *International Conference on Com- putational Intelligence and Communication Networks*, IEEE, India. 257–261. https://doi.org/10.1109/CICN.2014.66

[12] Dasgupta J., Bhattacharya K., & Chanda B., (2016). A holistic approach for off-line handwritten cursive word recognition using directional feature based on arnold transform, *Pattern Recognition Letters*, 79(1) 73-79. https://doi.org/10.1049/iet-ipr.2019.1398

[13] Shanjana, C., James, A., (2015). Offline recognition of malayalam handwritten text, *Procedia Technology*, 19,772–779. https://doi.org/10.1016/j.protcy.2015.02.109

[14] Alex, Meenu & Das, Smija, (2016). An approach towards malayalam handwriting recognition using dissimilar classifiers, *Procedia Technology*, 25, 224–231. https://doi.org/10.1016/j.protcy.2016.08.101

[15] Bag S., & Krishna A., (2015). Character segmentation of hindi unconstrained handwritten words, *International workshop on combinatorial image analysis*, Springer. 247–260. https://doi.org/10.1007/978-3-319-26145-4_18

[16] Barua S., Malakar S., Bhowmik S., Sarkar R., Nasipuri M., (2017). Bangla hand written city name recognition using gradient-based feature, *Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications*, Springer, 515, 343–352. https://doi.org/10.1007/978-981-10-3153-3_34

[17] Bhowmik S., Malakar S., Sarkar R., Basu S., Kundu M., & Nasipuri M., (2019). Off-line bangla handwritten word recognition: A holistic approach, *Neural Computing and Applications,* 31(10), 5783–5798. https://doi.org/10.1007/s00521-018-3389-1

[18] Pramanik R., Bag S., (2020). Segmentation-based recog- nition system for handwritten bangla and devana- gari words using conventional classification and trans- fer learning, *IET Image Processing*, 14(5), 959–972. https://doi.org/10.4018/IJCVIP.2017010104

[19] Tamen Z., Drias H., Boughaci D., An efficient multiple classifier system for arabic handwritten words recognition, *Pattern Recognition Letters*, 93(1), 123-132. https://doi.org/10.1016/j.patrec.2017.01.020

Funding:

No funding was received for conducting this study.

Conflict of interest:

The Authors have no conflicts of interest to declare that they are relevant to the content of this article.

About The License: