IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁝⁝ Vol. 11, Issue 2, February 2022

DOI: 10.17148/IJARCCE.2022.11236

Form Scanner & Decoder : Conversion of Text from any Application Form and its Language Translation Using OCR

Harish Kumar[1], Anshal Prasad[2], Ninad Rane[3], Nilay Tamane[4], Dr. Sharmila Sengupta[5]

UG-Computer Engineering, Vivekanand Education Society Institute of Technology, Mumbai[1-4]

Assistant Professor, Vivekanand Education Society Institute of Technology, Mumbai[5]

Abstract: Computers and phones may be more common than ever, but most people still prefer the traditional way of writing with ink on paper. People in the rural parts of India are mostly comfortable with the pen and paper way of going about their work. But with rapid technology advancements, everything has gone digital from Aadhar card forms to Birth Certificates. Despite this easy availability of a vast number of technical writing tools, many people choose to take their notes traditionally in the written manner in the language they are comfortable with, which is usually Hindi. Our work is on word recognition of handwritten Hindi characters and its implementation on handwritten forms. Our paper introduces an end-to-end word spotting system for the Hindi language using Segmentation based approaches. Our proposed architecture implements an end-to-end strategy that recognizes handwritten Hindi words from printed forms and is translated into English. Hence, handwriting recognition and translation interpret the Hindi handwritten input from various handwritten sources, such as paper documents, forms, into digital form translated into English. A form recognition system handles the formatting, performs correct segmentation into characters, and detects the Hindi words, which are then translated into English and shown on the form. The computational study of people's opinions, sentiments expressed is termed as sentiment analysis which is also known as opinion mining. For the feedback forms, sentiment analysis is performed using Random Forest algorithms and NLTK libraries like Porter stemmer and Stop words are used giving an accuracy of 88%.

Keywords : OCR , NLP , Image Processing , Sentiment analysis

I. INTRODUCTION

Hindi handwritten forms are not perfect and are difficult to understand. Handwriting recognition is considered one of the challenging areas of research when it comes to the field of pattern recognition and image processing. This contributes immensely to the development of the automation process in India and enhances the interface between man and machine in numerous applications. Several research works have been focusing on new ways that would reduce the pre-processing time for detecting and extracting the text while providing higher recognition accuracy but they are for other languages and not in Hindi. Handwriting recognition is a challenging task for a large number of reasons. The primary reason is that different people have different styles of writing. The secondary reason is there is a wide range of characters like capital letters, small letters, digits, and special symbols. Optical Character Recognition (OCR) is a technology developed to recognize characters in images of printed documents. These types of technologies are not directly applicable for handwritten images due to various types of challenges like the different styles of handwriting, presence of noise and distortions in them, and blur which is often present in natural images or pictures. To overcome this problem of handwritten-based search in documents, finding the relevant information is proposed as an alternative. This type of system is called a word spotting system as it tries to spot occurrences of words on a document page. The spotting of words appears under two distinct trends wherein the fundamental difference is regarding the search space which can be a set of segmented word images or the complete document image. In this work, we look at the word spotting problem in a segmentation-based multi-writer scenario. The Hindi language is used widely throughout India. It consists of 10 numerals(०, १, २, ३, ४, ५, ६, ७, ८, ९) and 36 consonants (क, ख, ग, घ, ङ, च, छ, ज, झ, ञ, ट, ठ, ड, ढ, ण, त, थ, द, ध, न, प, फ, ब, भ, म, य, र, ल, व, श, ष, स, ह, क्ष, त्र, ज्ञ). Some consonants are complex and made by combining some others. This dataset is trained and is used as our word spotting system for recognizing the handwritten letters on the form and later converting them into English. We created our own version of the dataset where the format of the image was Grayscale with 2 pixels margin on each side. Sentiment analysis measures the inclination of the opinions of various people through natural language processing (NLP), computational linguistics and text analysis, which are used to extract and analyze information. The analyzed data summarizes the general public's reactions toward certain products, people or ideas and reveal the contextual

DOI: 10.17148/IJARCCE.2022.11236

meaning of the information. The feedback forms having the computer encoded english text is then compared with our dataset having opinions using NLP.

II . LITERARY REVIEW

Research on Text Detection and Recognition Based on Recognition Technology Authors: Yumming He Description: This paper proposes a system of image classification, detection of objects, and semantic segmentation using Deep Learning. It is a data-driven algorithm that makes use of vast datasets which is time-consuming and leads to having a lower accuracy rate of detection. This paper does not include translation of the recognized text. This system must be trained about object information. Feature extraction is also a part of the process to be improved here.

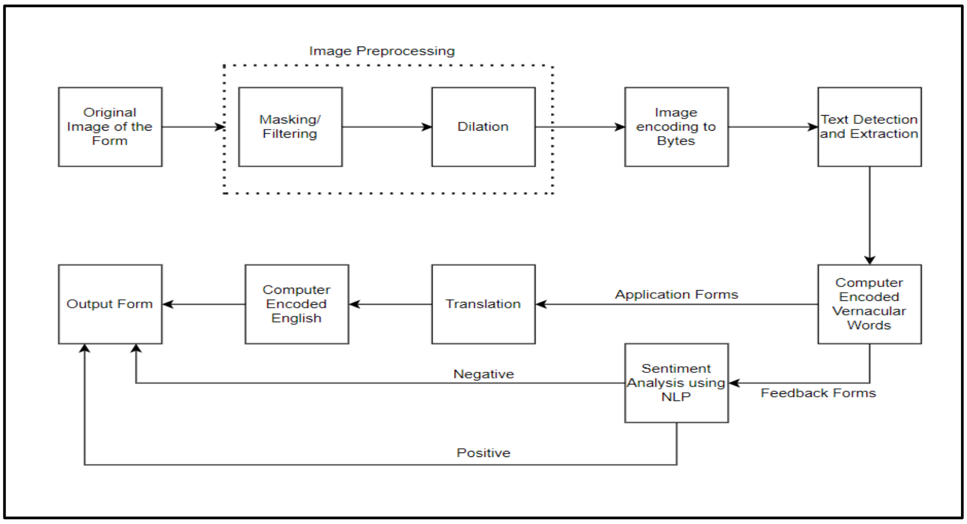
Towards accurate scene text recognition with Semantic Reasoning Networks Authors: Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, Rui Ding Description: This paper proposes a system where a trainable framework named semantic reasoning network (SRN) is used for scene text recognition, where a global semantic reasoning module (GSRM) captures the semantic context through multi-way parallel transmission. These results are shown on 7 public benchmarks, including regular text, irregular text, and non-Latin long text to verify the effectiveness and robustness of the proposed method. This paper looks for the recognition and detection of Latin text along with its translation into English using various parameters, though the accuracy is relatively low here.

Deep adaptive learning for writer identification based on handwritten word images Authors: He, Sheng; Schomaker Description: This paper proposes a new adaptive convolutional layer to make use of the learned deep features. A neural network having one or more adaptive convolutional layers is trained end-to-end so as to exploit robust generic features for a specific task. Lexical Content, Word length, and Character attributes are the three auxiliary tasks that correspond to three explicit features of handwritten word images. The recognition of letters and their translation here becomes time-consuming due to the algorithm

RNN based handwritten word recognition in Devanagari and Bengali scripts using horizontal zoning: Rajib Ghosh, Chirumavila Vamshi, Prabhat Kumar Description: This paper has implemented a convolution neural network in many ways, which include - i] Training the CNN model from scratch in a sequential manner. ii] Using MobileNet: Transfer the learning paradigm from a pre-trained model on a Tamizhi dataset. iii] Building the model with CNN and SVM. iv] SVM has the best accuracy for the recognition of handwritten Bangla characters.

III. PROPOSED SYSTEM

System Architecture :

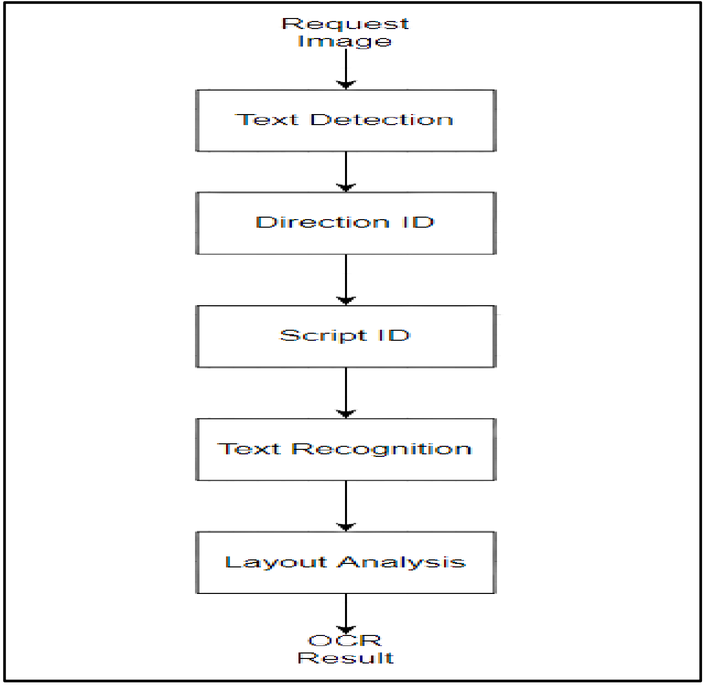


Images from the input folder location are allocated into a list, then the image is given an image variable. The word spotter detects a text function that calculates the coordinates of the handwritten text. The preprocessing is done on the input image and Masking/filtering is done using the lower and upper HSV values. After masking, the process of encoding is done where the image is converted into strings of data. The Image is filtered where dilation and erosion processes are done. Filtering is carried out in order to filter out the black from the image. Bounding boxes around the texts are removed after finding out their areas, after which dilation, erosion for enhancement. The image is then encoded into strings of data. After this, only blue ink text is present. For replacement of this handwritten text, the word spotter gives the

DOI: 10.17148/IJARCCE.2022.11236

coordinates around the blue ink text. Text extraction is done on this, which is translated and added into a new list. We now have bounding boxes and the translated text. The original image is now thresholded, then using the coordinates, we white out the handwritten text and replace it with the words from the list present. The image of the input handwritten form is taken after which the necessary pre-processing is done on it. Further, Dilation and Erosion processes are performed on it. Using Dilation, Pixels can be added to the boundaries of objects in an image whereas, with the help of Erosion, Pixels are removed on object boundaries. The number of pixels added or removed from the objects in an image is dependent on the size and shape of the structuring element that is used to process the image. During the morphological dilation and erosion operations, the state of any given pixel in the output image is determined by applying a rule to the corresponding pixel and its neighbors in the input image. Image Encoding into Bytes is done in the next stage after which the handwritten text is recognized. The recognized text is then extracted for its translation after which an output image gets generated where the handwritten Hindi text is recognized and translated into English. After the translation of the handwritten text into English, the output image of the form is generated with the handwritten part of the form replaced with computerized English text present. Coming to the feedback forms, sentiment analysis is performed on them using Random Forest algorithms and NLTK libraries like Porter stemmer and Stop words are used giving an accuracy of 88%.

OCR (Optical Character Recognition) -



- Text detection - Uses CNN to locate sentences and create bounding boxes. By using multi-layer CNNs, powerful text detection modules are trained. CNN is trained with images having marked regions to increase the accuracy. CNN is useful to process low-level features and high-level context.
- Direction ID - Classifies the directions for each bounding box.
- Script ID - Locates the script in the bounding box. It can allow multiple scripts per image but initially, only 1 script per bounding box is assumed.
- Text recognition - This is the most important part of optical character recognition in which all the text part is identified from an image. CNN recognizes the alphabets and other characters by finding the contrast in their features. A CNN classifier architecture has a number of convolutional layers which perform the feature extraction and fully connected layers, followed by a soft-max layer for classification.
- Layout analysis - This determines the reading of order and distinguishes the title and headers.

CONVOLUTIONAL NEURAL NETWORKS

CNN consists of a huge number of neurons that are interconnected and weights and biases. Neurons are organized as layers in this structure. CNN contains different layers like the input layer, output layer, and various hidden layers. The neural network is called a deep neural network if it has a large number of hidden layers. Neurons in the hidden layer are connected to the receptive field of input space generated from the previous layer. Thus, the connection weights are reduced because of this approach. Hence, the training time for CNN is less. 2D arrays of data are the inputs for a generic

CNN. The CNN layers are arranged in three dimensions, the three being width, height, and depth. Following is the basic information about the layers -

- Input layer - It acts as a buffer to hold the input and also to pass it to the subsequent layers.
- Convolution Layer - The feature extraction operation happens here. The process of convolution operation involves sliding the kernel over input and also performing the sum of products. The size of the step with which the kernel will slide is called stride. Multiple different feature maps are made by various convolution operations which are done by kernels on the input. The depth of the layer would be the number of feature maps.
- Rectified Linear Unit - This brings about non-linearity. To fasten the process of learning, all values less than zero are replaced by zero. Outputs generated by the previous layer are passed through the activation function.
- Pooling layer - This layer reduces the spatial size of every feature map. This, in turn, lowers the computation. This layer also makes use of a sliding window that moves across the feature map and transforms them into values.
- Fully connected layer - This layer connects all neurons of the previous layer to each neuron in the current layer. It helps in classifying the output. The arrangement of layers is different for classification and regression problems. For a regression problem, a fully connected layer is followed by a regression layer for output prediction. For the classification problem, the next layer is the soft-max layer which helps to get the probability of every class.

Dataset: The dataset used for building the OCR model is DHCD (Devanagari Handwritten Character Dataset) of handwritten digits having a training set of 78200 letters, and a test set of 13800 letters for 46 characters from क to ज्ञ as alphabets and ० to ९ as numerals. The characters are written on paper first, which are later scanned and cropped manually. Some screenshots of the characters are shown below.

Character 'क' :

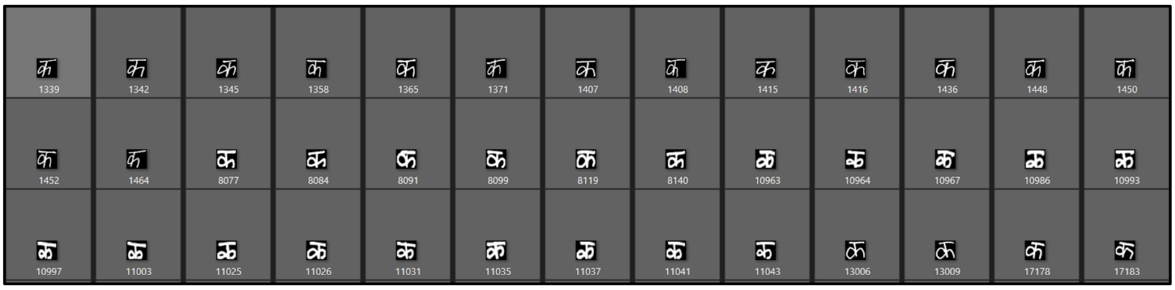


Figure 1

Character 'ख' :

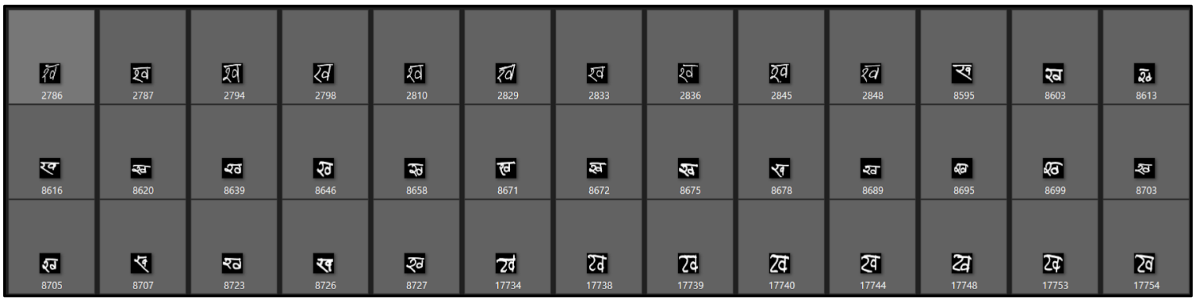


Figure 2

Character 'ग' :

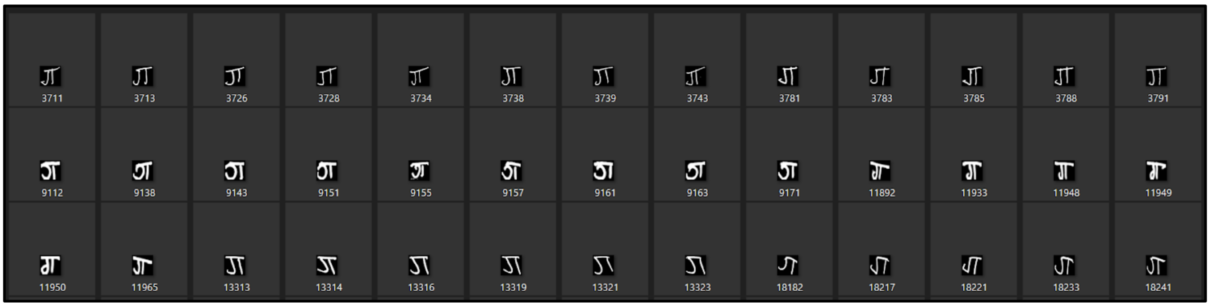
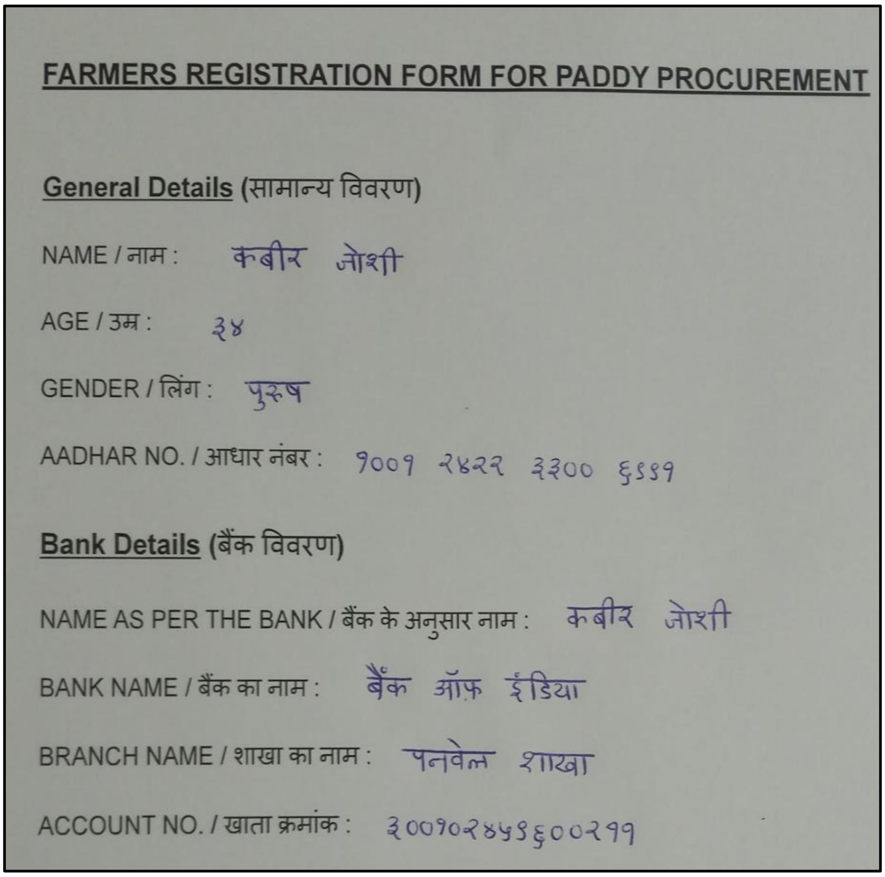


Figure 3

DOI: 10.17148/IJARCCE.2022.11236

Input image :



FARMERS REGISTRATION FORM FOR PADDY PROCUREMENT

General Details (सामान्य विवरण)

NAME / नाम : कबीर जोशी

AGE / उम्र : ३४

GENDER / लिंग : पुरुष

AADHAR NO. / आधार नंबर : ९००१ २४२२ ३३०० ६५५१

Bank Details (बैंक विवरण)

NAME AS PER THE BANK / बैंक के अनुसार नाम : कबीर जोशी

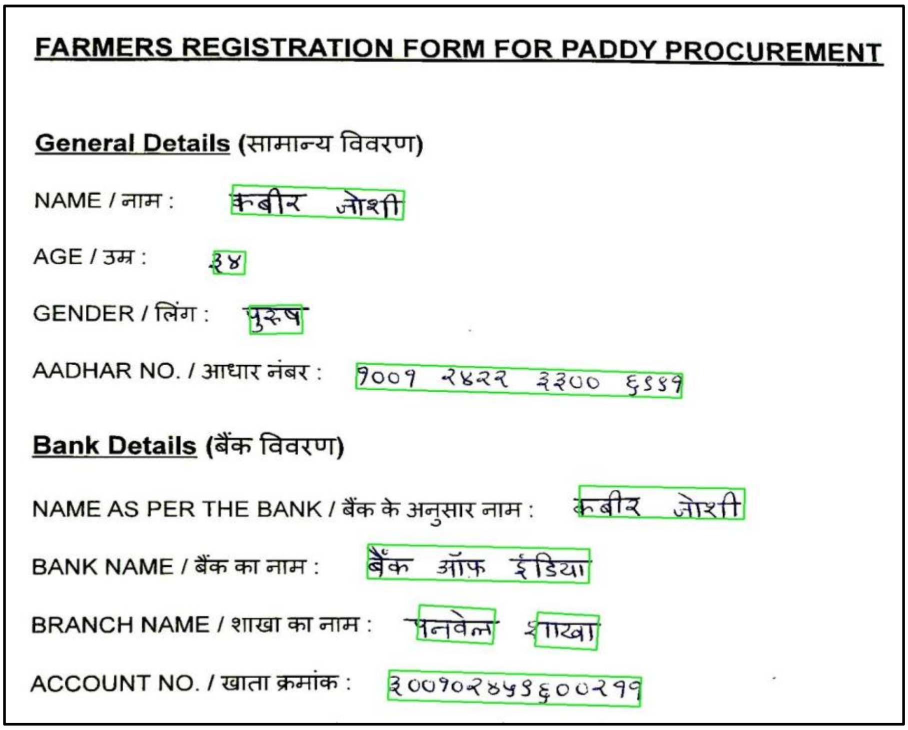
BANK NAME / बैंक का नाम : बैंक ऑफ इंडिया

BRANCH NAME / शाखा का नाम : पनवेल शाखा

ACCOUNT NO. / खाता क्रमांक : ३००१०२४५५६००२११

Input image of the form where it is filled in Hindi. An image is taken and uploaded as the input which is given an image variable.

Bounding Boxes :



FARMERS REGISTRATION FORM FOR PADDY PROCUREMENT

General Details (सामान्य विवरण)

NAME / नाम : कबीर जोशी

AGE / उम्र : ३४

GENDER / लिंग : पुरुष

AADHAR NO. / आधार नंबर : ९००१ २४२२ ३३०० ६५५१

Bank Details (बैंक विवरण)

NAME AS PER THE BANK / बैंक के अनुसार नाम : कबीर जोशी

BANK NAME / बैंक का नाम : बैंक ऑफ इंडिया

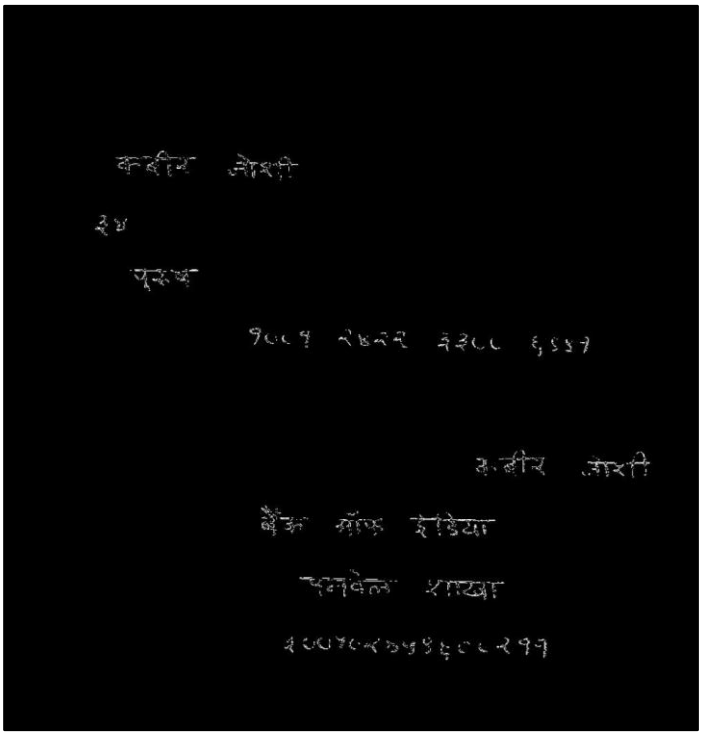
BRANCH NAME / शाखा का नाम : पनवेल शाखा

ACCOUNT NO. / खाता क्रमांक : ३००१०२४५५६००२११

DOI: 10.17148/IJARCCE.2022.11236

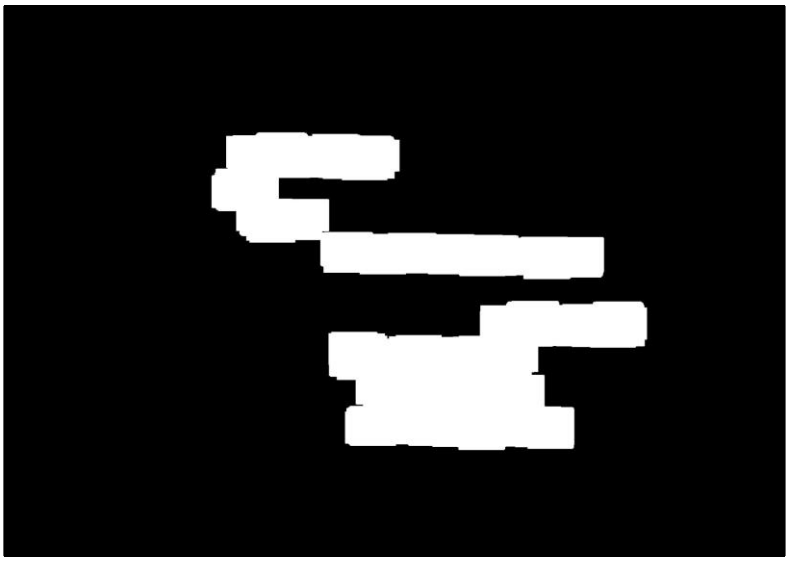
The input image is then thresholded. With the help of a list of coordinates calculated by our OCR and using the CV2 line method, the handwritten input sentences are boxed. The thresholded image is now in black and white and the bounding boxes are present around the words.

Masked image :



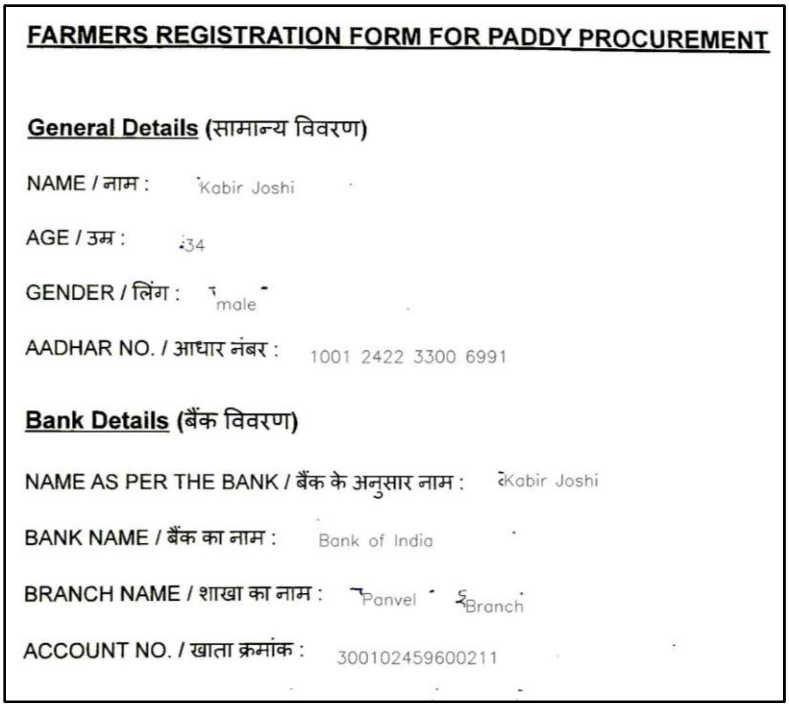
The rest of the area is then masked apart from the bounding boxes. This process of masking is done by calculating the upper and lower HSV values.

Dilated Image :



Using Dilation, Pixels can be added to the boundaries of objects in an image. The number of pixels added from the objects in an image is dependent on the size and shape of the structuring element that is used to process the image. During the morphological dilation operations, the state of any given pixel in the output image is determined by applying a rule to the corresponding pixel and its neighbors in the input image. After dilation, the process of encoding is done where the image is converted into strings of data.

Output image :



FARMERS REGISTRATION FORM FOR PADDY PROCUREMENT

General Details (सामान्य विवरण)

NAME / नाम : Kabir Joshi

AGE / उम्र : 34

GENDER / लिंग : male

AADHAR NO. / आधार नंबर : 1001 2422 3300 6991

Bank Details (बैंक विवरण)

NAME AS PER THE BANK / बैंक के अनुसार नाम : Kabir Joshi

BANK NAME / बैंक का नाम : Bank of India

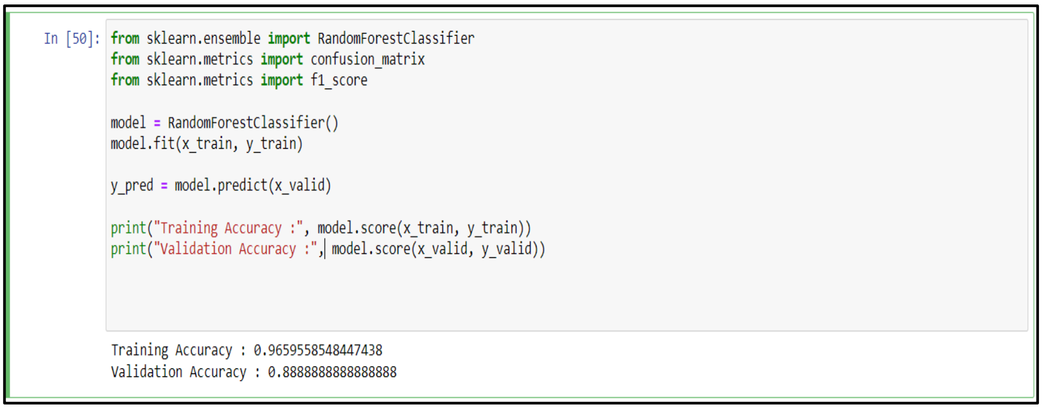
BRANCH NAME / शाखा का नाम : Panvel Branch

ACCOUNT NO. / खाता क्रमांक : 300102459600211

This is the output image where the handwritten Hindi text has been converted into computer encoded English text and has been overlaid in the same areas of the thresholded image. Translating a handwritten text from one language into another has turned out to be a fundamental task and the python package known as py-translate has helped to accomplish this task. It provides translation into standard language for a variety of input languages. Here Py-Translate converts the computer encoded Hindi text to computer encoded English text.

Sentiment Analysis using NLP

Sentiment analysis is the measure for inclination of the opinions of various people through natural language processing (NLP), computational linguistics and text analysis, which are used to extract and analyze information. The analyzed data summarizes the general public's reactions toward certain products, people or ideas and reveal the contextual meaning of the information. For the feedback forms, sentiment analysis is performed using Random Forest algorithms and NLTK libraries like Porter stemmer and Stopwords are used giving an accuracy of 88%.The computational study of people's opinions, sentiments expressed is termed as sentiment analysis which is also known as opinion mining.



```
In [50]: from sklearn.ensemble import RandomForestClassifier
         from sklearn.metrics import confusion_matrix
         from sklearn.metrics import f1_score

         model = RandomForestClassifier()
         model.fit(x_train, y_train)

         y_pred = model.predict(x_valid)

         print("Training Accuracy :", model.score(x_train, y_train))
         print("Validation Accuracy :", model.score(x_valid, y_valid))

         Training Accuracy : 0.9659558548447438
         Validation Accuracy : 0.8888888888888888
```

The feedback forms are taken for sentiment analysis where we have created a dataset of good and bad words. The English encoded words on the output are analyzed and compared with the two databases where sentiment analysis is done and the output is given as a positive or a negative review. Our current model uses the random forest algorithm where the dataset is trained and then tested.

DOI: 10.17148/IJARCCE.2022.11236

Random Forest is a very popular machine learning formula that belongs to the supervised learning technique. It is used for each Classification and Regression problems in Machine Learning. It supports the thought of ensemble learning, that is a method of combining multiple classifiers to reach the solution of a complex problem and to enhance the performance of the model. Random Forest is a classifier that contains a wide range of decision trees on a large number of subsets of the given dataset and takes the typical to enhance the predictive accuracy of that dataset. Rather than depending on one decision tree, the random forest takes the prediction from every tree and supports the maximum votes of predictions, and then predicts the ultimate output. The bigger variety of trees within the forest ends up in higher accuracy and prevents the matter of overfitting.

Library used -
NLTK (Natural Language Toolkit) - It is considered as one of the most powerful NLP libraries. It consists of packages which make machines understand human language and respond appropriately.
PorterStemmer() is a module in NLTK which implements Porter. Porter Stemmer is the original stemmer, known for its easy use and speed. The resultant stem is a shorter word with the same root meaning. Five steps of word reduction methods are used, each having its own set of mapping rules.
Stop words - Stop Words: A stop word is a commonly used word (such as “hi”, “them”, “these”, “out”) that a search engine has been programmed to ignore while indexing entries for searching and when retrieving them as the result of a search query.
We do not want these words to take up space in the dataset, or take up the processing time. Hence we can remove them easily, by storing a list of words that you consider to stop words.

Experimental results-
On testing different styles of handwriting in Hindi, the accuracy of the system detecting the coordinates of the handwritten text and replacing it with the English text in the same position was seen to be 60%. With the increase in noise and distortion, the accuracy of detection decreased, but with generic clean handwriting, the Hindi text is detected, translated, and replaced in the same spot with ease.

CONCLUSION-

This software will benefit a lot of people from rural areas with an easy way to fill their forms without having to worry about the language. They would just have to upload an image of the form filled in their native language on the app. The software would then convert the text written in the native language into English and the new form would be returned back to the person. The main steps in the working of this app would be text detection, text recognition, and translation. This would decrease the dependency of such people on others and would make the process of form filling hassle-free. We also implemented sentiment analysis on feedback forms using random forest and NLP libraries and achieved an accuracy of 88%.

REFERENCES

1. https://stacks.stanford.edu/file/druid:my512gb2187/Ma_Lin_Zhang_Mobile_text_recognition_and_translation.pdf
2. https://ijesc.org/upload/9086956de6da268b03fdaece6a3f6cd7.Text%20Detection%20and%20Translation.pdf
3. https://www.ijert.org/research/machine-learning-approach-for-translating-handwritten-document-to-digital-form-IJERTCONV9IS02003.pdf
4. http://norvig.com/spell-correct.html
5. https://www.researchgate.net/publication/331590226_A_Study_on_Text_Recognition_using_Image_Processing_with_Datamining_Techniques
6. https://www.researchgate.net/publication/298808334_Handwritten_Text_Recognition_System based_on_Neural_Network
7. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9151144
8. V. V. Mainkar, J. A. Katkar, A. B. Update and P. R. Pednekar, "Handwritten Character Recognition to Obtain Editable Text," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 599-602, doi: 10.1109/ICESC48915.2020.9155786.
9. Sabeenian R.S., Vidhya M. (2010) Handwritten Text to Digital Text Conversion using Radon Transform and Back Propagation Network (RTBPN). In: Das V.V., Vijaykumar R. (eds) Information and Communication Technologies. ICT 2010. Communications in Computer and Information Science, vol 101. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-15766-0_82
10. Kumar M., Jindal M.K., Sharma R.K. (2011) Review on OCR for Handwritten Indian Scripts Character Recognition. In: Nagamalai D., Renault E., Dhanuskodi M. (eds) Advances in Digital Image Processing and Information

Technology. DPPR 2011. Communications in Computer and Information Science, vol 205. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-24055-3_28

11. **MDPI and ACS Style**
12. Albahli, S.; Alhassan, F.; Albattah, W.; Khan, R.U. Handwritten Digit Recognition: Hyperparameters-Based Analysis. Appl. Sci. **2020**, 10, 5988. https://doi.org/10.3390/app1017598
13. P. Mishra, P. Pai, M. Patel and R. Sonkusare, "Extraction of Information from Handwriting using Optical Character recognition and Neural Networks," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1328-1333, doi: 10.1109/ICECA49313.2020.9297418.
14. R. R. Ingle, Y. Fujii, T. Deselaers, J. Baccash and A. C. Popat, "A Scalable Handwritten Text Recognition System," 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 17-24, doi: 10.1109/ICDAR.2019.00013.
15. S. A. Ayyadevara, P. N. V. S. R. Teja and M. Rajesh Kumar, "Handwritten Character Recognition Using Unique Feature Extraction Technique," 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2018, pp. 1239-1243, doi: 10.1109/RTEICT42901.2018.9012248.
16. J. J. Hull, "A database for handwritten text recognition research," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 16, no. 5, pp. 550-554, May 1994, doi: 10.1109/34.291440.