Text Recognition from Scanned Documents

Sujata Desai, Veena Patil, Deepa Dasu Pawar

Abstract— **Digitalization is the one of the trend in today's world, it helps us to do our work more quickly and efficiently. Optical Character Recognition (OCR) is the one of the technique used for recognition of text from scanned documents in digitalization. In this work, a method has been implemented for recognition of English handwritten text from scanned documents. Here zone based feature extraction technique, support vector machine and HOG features are proposed. This method contains some steps like preprocessing, segmentation, feature extraction and classification. The proposed method can recognition and verify the character with the maximum accuracy rate.**

Index Terms— **HOG features, Scanned documents, support vector machine (SVM), zoning.**

I. INTRODUCTION

Optical Character Recognition (OCR) is the area of interest, where a lot of research is going on this area due to presences of its complexity in this area, the problem not yet solved. It is the one of the technique deals with the recognition of handwritten characters in it and input can served directly to the computer. There are many different approaches to perform character recognition. One of the approach is performed based on HOG features, support vector machine (SVM) and zone based techniques. Some steps are important for the development of reliable handwritten character recognition, the first step is extraction of features from handwritten images and the second method is classification of new character images.

Dataset of handwritten compound character is not yet available for English handwritten text and hence we created our own dataset for compound character as shown in the figure 1 and these data set are examined with the proposed method. The database has been created for English handwritten text, it contains split character. These dataset are created by different volunteers, and it in the form of Bitmap images.

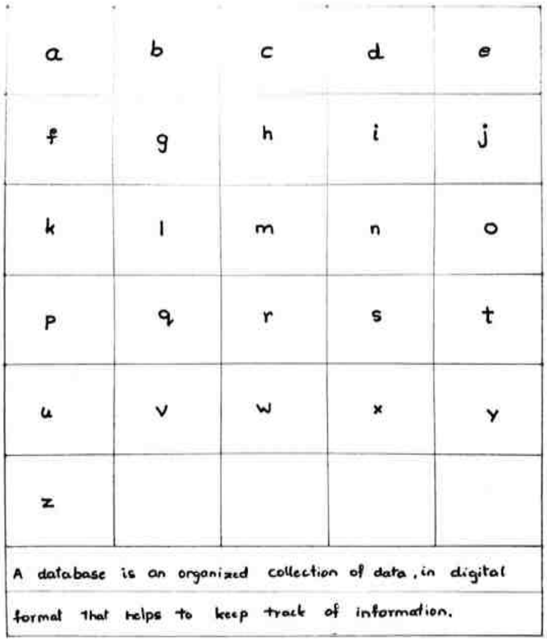
Manuscript received May 26, 2019

Sujata Desai, Department of Computer Science and Engineering, BLDEA'S V.P. Dr. P. G. Halakatti College of Engineering and Technology, Vijayapur – 586103, India, 9448933005,
(*e-mail:* shreeja.clg@gmail.com)

Veena Patil, Department of Computer Science and Engineering, BLDEA'S V.P. Dr. P. G. Halakatti College of Engineering and Technology Vijayapur – 586103, India, 9886034571,
(*e-mail:* veenaanandapatil@gmail.com)

Deepa Dasu Pawar, Department of Computer Science and Engineering, BLDEA'S V.P. Dr. P. G. Halakatti College of Engineering and TechnologyVijayapur – 586103, India, 9008404295,
(*e-mail:* deepadasupawar20@gmail.com).

The main objective is to develop an OCR for handwritten English text, the input to the system would be a pure handwritten English text and output would be a recognized English characters.



a	b	c	d	e
f	g	h	i	j
k	l	m	n	o
p	q	r	s	t
u	v	w	x	y
z				

A database is an organized collection of data, in digital
format that helps to keep track of information.

Fig. 1. Dataset

The paper is organized as follows, Section I contains the introduction about techniques used for recognition purpose, Section II contains related work reviews, Section III explains proposed methodology with basic diagram, Section IV describes results and discussion about proposed method, and finally conclusion is discussed in Section V.

II. RELATED WORK

Mr. Pratik Madhukar Manwatkar et al. [1] the information about paper documents can be stored as scanned documents in a computer in a computer system, but it is difficult to reuse these information. Due to the complexity, that is font characteristics of the character while reading them. To overcome from these challenges, the Author proposed a method which is divided into four modules i.e. preprocessing, system training, text recognition and post processing modules. Here matrix feature extraction method and Kohonen neural network methods involved for the feature extraction and classification purpose.

Rafael Dveire lins et al. [2] Binarization method support for all documents but it does not support for complex documents such as encompass photos, charts and texts so on. To overcome from this disadvantage, the Author proposed a new concept, here image processing depends on nature of the elements. The new concept involves block spotting, block classification, block Binarization and document reassembling. Here documents are divided into blocks and these individual block are classified, once done with classification these block were recognized and binarized, Binarization of blocks done directly using Otsu algorithm. Then reassemble of image done and known as monochromatic image, it should be less than half of the size then the original image.

Banumathi K. L. et al. [3] here segmentation of text documents can be done using projection profile method, it is a one of the phase for the recognition of contents. The proposed method for line segmentation provides 90% of segmentation result, which can be done using horizontal projection profile, but segmentation of word needs more effort due to lack of space between the text lines and overlapping of subscripts there is decrease in accuracy.

Narasimha Reddy Soora et al. [4] manual search can be done for huge number of scanned Indian documents by using robust automatic searching software. Work is going on towards indexing of aged printed multingual Indian office documents. The geometrical technique is proposed by the Author, which group the components having multi orientations and local skew. The proposed method was evaluated by having English Devanagari, and Marati scripts. Line segmentation achieves 99% of accuracy. In Devanagari scripts word in one line is close to the top of a word in the text line, due to this drawback there is drop in accuracy rate.

III. Methodology

The proposed system it deals with some of the steps in digital image processing i.e. data acquisition, pre-processing, segmentation, feature extraction, and classification as shown in the below figure 2:

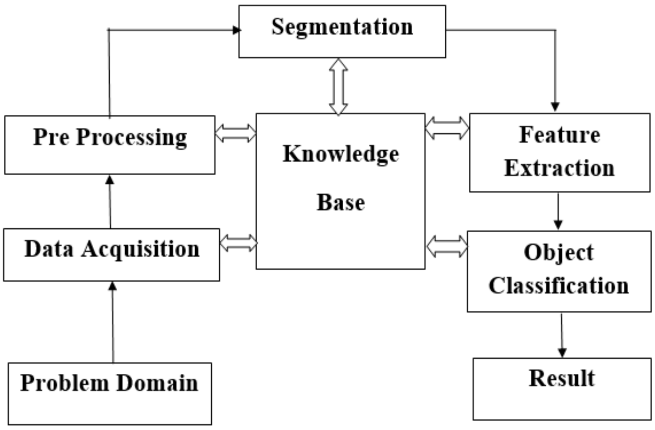


Fig. 2. Steps involved in proposed methodology

Steps involved in methodology listed below:

1. Data acquisition, it acquires the required data for the working purpose.
2. Preprocessing, it involves two method in it i.e. noise removal and Binarization.
3. In noise removal, it remove the noise and smoothen the image for any pixels.
4. In Binarization, conversion of gray scale image into binary image takes place with a proper threshold assigning black pixels to data and white pixels to background.
5. In segmentation, samples collected are words in discontinuous format, these character are separated by bounding lines, segmentation of word done by calculating horizontal projection profile technique.
6. In feature extraction, it extracts information of interest and extraction of image done by using zone based technology, and image dilation.
7. Classification, which build classification model from an input data by using Support Vector Machine (SVM) techniques.

A. Data Acquisition

This is the step in image processing application, where data required is acquired in the image format as shown in the above figure 3:

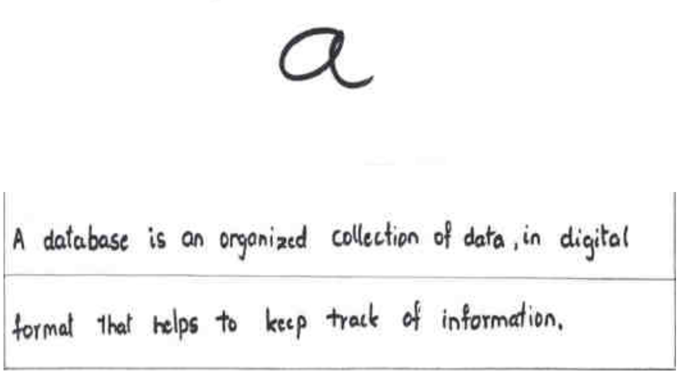

a
A database is an organized collection of data, in digital
format that helps to keep track of information.

Fig. 3. The Original Image.

B. Preprocessing

It involves noise removal and Binarization technique within it.

- Noise removal:
  Here 3×3 neighborhood averaging filter is used to smoothen the image for any noisy pixels.

- Binarization:

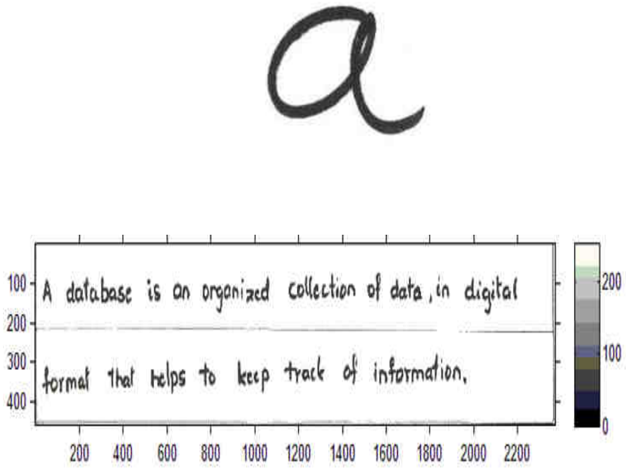


Fig. 4. The Gray scale Image.

It converts the grayscale image as in the above figure 4. into binarized image, here Binarization of image done with a proper threshold assigning black pixels to data and white pixels to background as in the below figure 5.

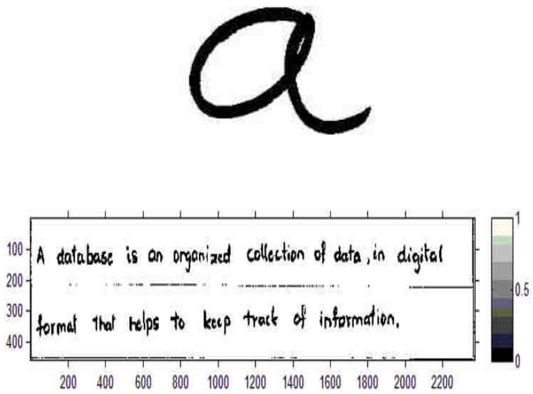

A database is an organized collection of data, in digital
format that helps to keep track of information.
100
200
300
400
200
400
600
800
1000
1200
1400
1600
1800
2000
2200
1
0.5
0

Fig. 5. The Binarized Image.

C. Segmentation

Segmentation of an image is partitioned into its parts, here collection of samples are in discontinuous form. Characters are separated by bounding line and word is segmented into characters by calculating horizontal projection profile.

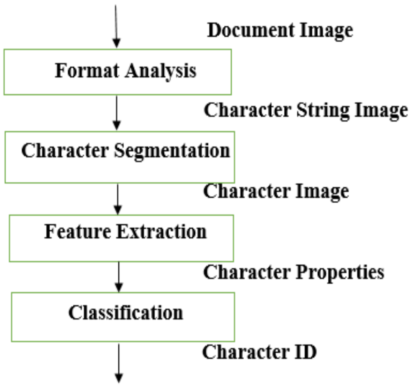


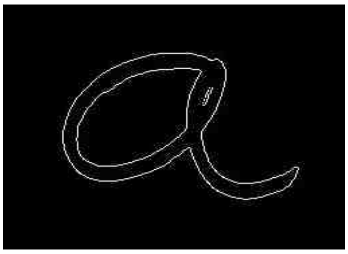
Fig. 6. The analysis of Segmentation.

The above figure 6 shows the analysis of segmentation, the sequence of segmentation is repeated until no additional character images are found.

D. Feature extraction

It deals with extraction of feature from images. Here scale invariant based feature extraction algorithm implemented here. The character has specific features which are extracted from the samples. Some of technique explained, which is used in Scale Invariant feature based recognition that is as follows:

- Image Dilation:



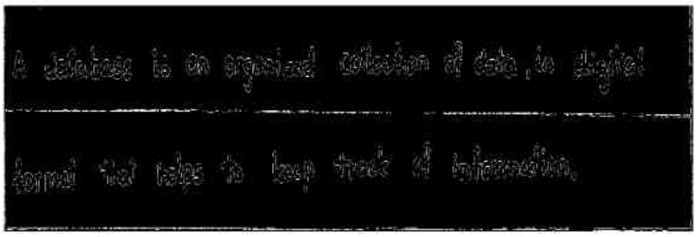

A database is an organized collection of data, in digital
format that helps to keep track of information.

Fig. 7. Edge detected Image.

Dilation is developed for binary images, it is one of the basic operation involved in mathematical morphology. It basically uses a structuring element for probing and expanding the shape contained in the input, first it expanded to gray scale and then to complete lattices.

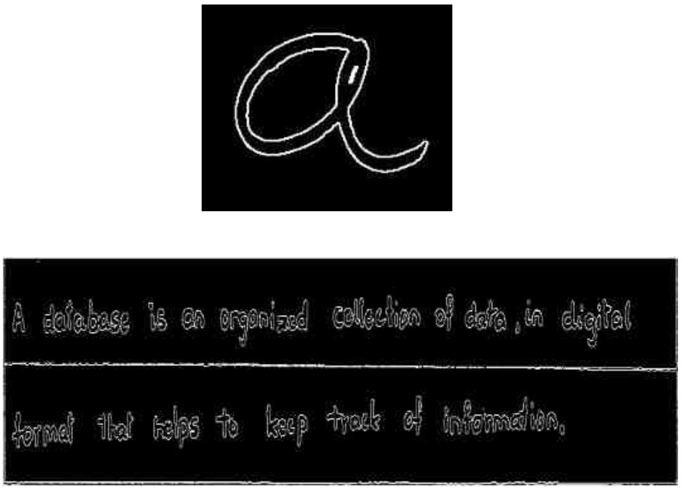

A database is an organized collection of data, in digital
format that helps to keep track of information.

Fig. 8. Edge dilated Image.

- Zoning:

After the edge dilation of the image is done, the image is divided into regions of equal size that is zoned into 9 equal sized regions. Proposed method was applied to individual region rather than whole image, so that we get fine information about each regions.

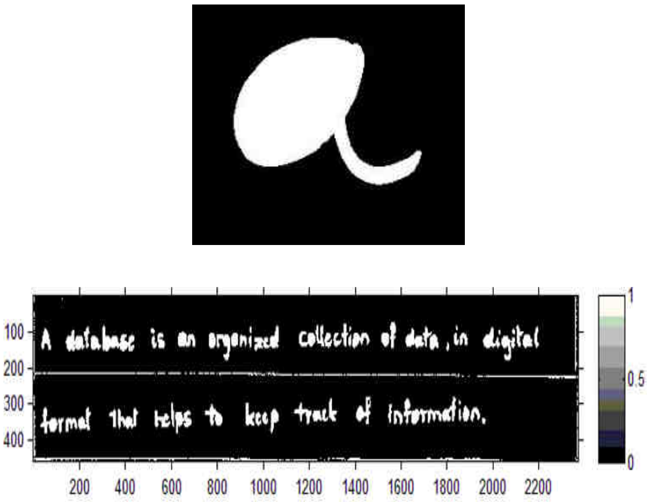

A database is an organized collection of data, in digital
format that helps to keep track of information.
100
200
300
400
200
400
600
800
1000
1200
1400
1600
1800
2000
2200
1
0.5
0

Fig. 9. The feature extracted images.

E. Classification

Classification of an object is one of the task involved in computer vision application, here Support Vector Machine based classifier is used. The basic steps for creating an object classifier is listed below:

- Select the data set with labeled images.
- Dividing the data set into two sets i.e. training set and testing set.
- Test the classifier using feature extracted image from training set and testing set.

Here evaluation of character can be done using images from the testing set and accuracy for the classifier can be generated using a confusion matrix. For this evaluation, the character ‘a’ is misclassified as character ‘d’ and similarly character ‘s’ is misclassified as character ‘r’ most likely due to their similar shapes.

- HOG features

For classification purpose HOG feature vectors are used, it is important to encode amount of information about object. The ‘extractHOGFeatures’ function which returns an output. The effect of feature vector can be represented as cell size parameter, it has amount of shape information encoded as shown in figure 10:

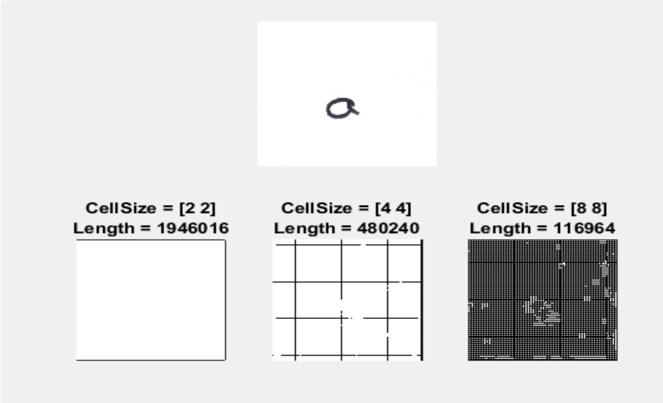


Fig. 10. Shows the effect of the cell size parameter.

F. Knowledge Base

Knowledge base is a region of an image where information of interest is known to be located. It is used in the classification stage to compare the test sample feature vector with those of all writers enrolled.

IV. Results and Discussion

Experiments were conducted on 2 types of set, training data set and testing data set. The results obtained for these data set are displayed by using confusion matrix and it displayed in the form of table respectively and displays maximum rate of accuracy as result.

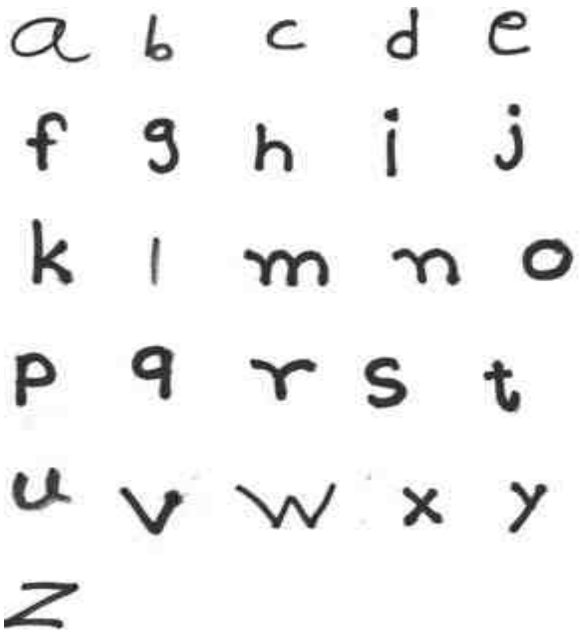


Fig. 11. The gray scale image

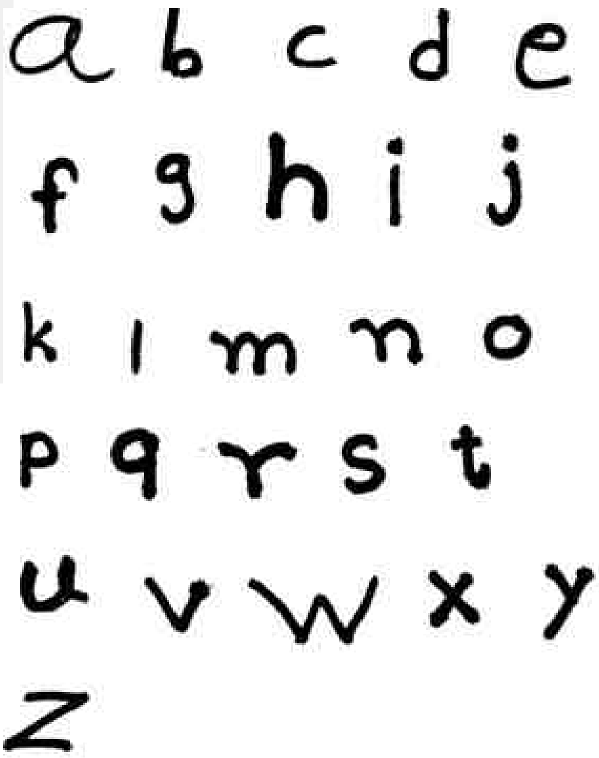


Fig. 12. The binarized image.

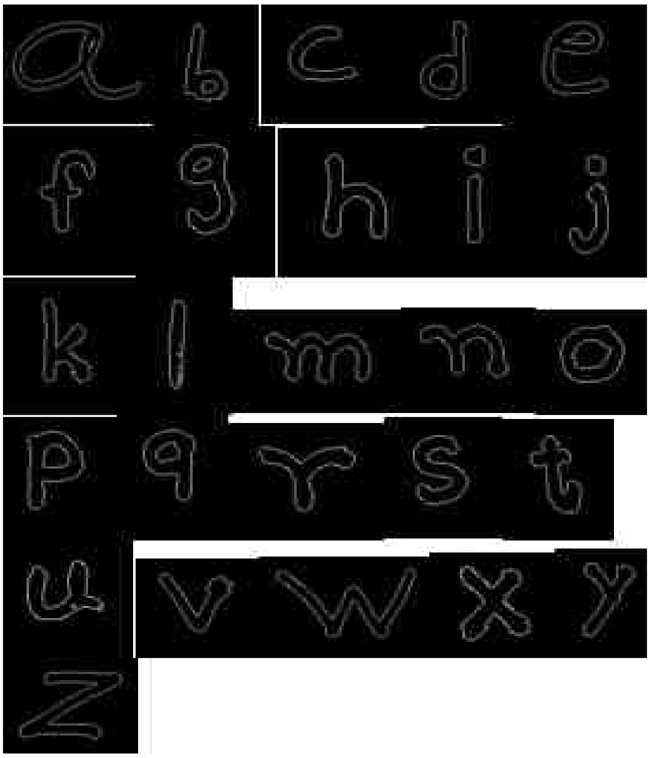


Fig. 13. The edge detected image.

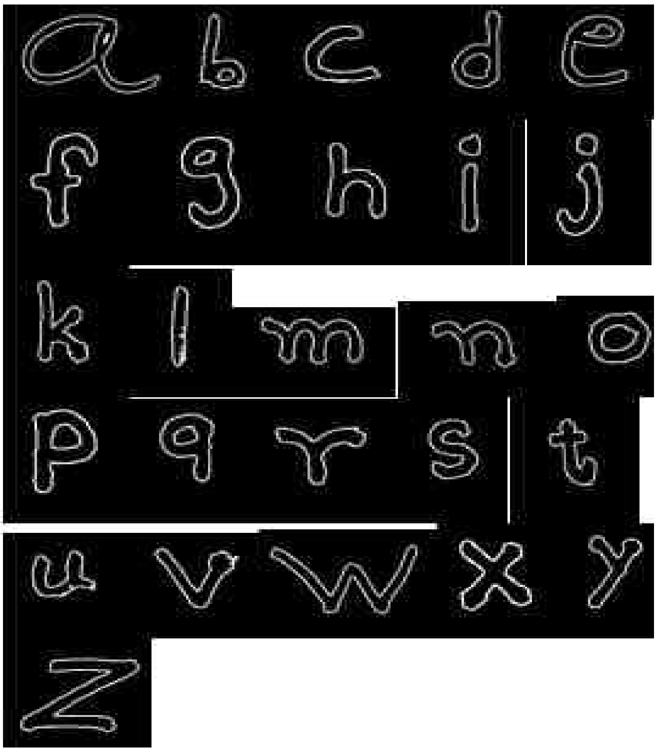


Fig. 14. The edge dilated image.

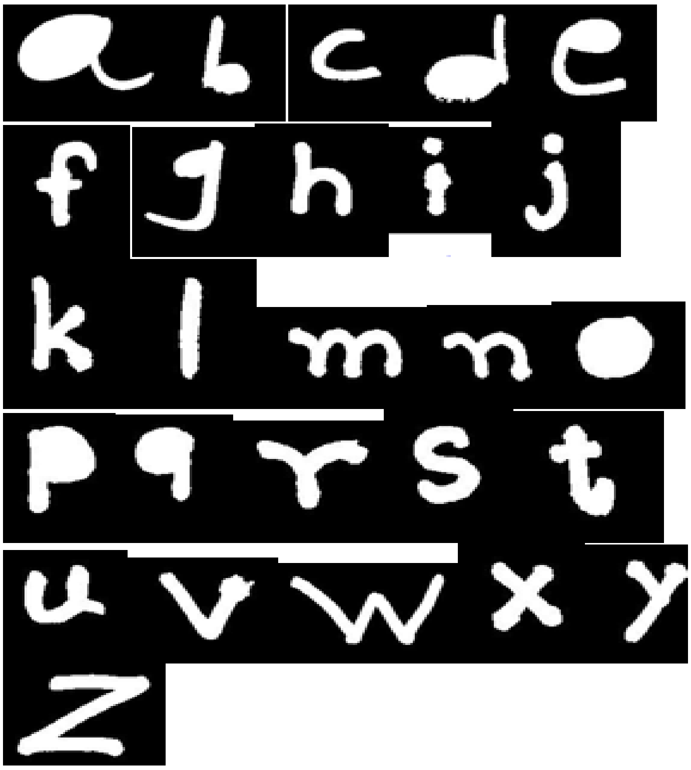


Fig. 15. The extracted image

V. Conclusion

The proposed work is to recognize the English handwritten text images, it involves scale invariant based feature recognition that is HOG features and Support Vector Machine based classifier that is confusion matrix. Proposed method involves 2 types of dataset, training set and testing set to get the confusion matrix. As a result every character is recognized and over all maximum accuracy is achieved. Due to the presence of noise in the image and dissimilar identification of character, there is a drop in accuracy rate.

REFERENCES

[1] Mr. Pratik Madhukar Manwatkar and Mr. Shashank H. Yadav, “Text Recognition from Images”, IEEE, 2015.

[2] Rafael Dveire Lins, Gabriel de Franca, P. eSilva and Marcos Martins de Almeida, “Binarizing Complex Scanned Documents”, IEEE, 2015, pp. 56-60.

[3] Banumathi K. L and Jagadeesh Chandra A. P, “Line and Word Segmentation of Kannada Handwritten Text documents using Projection Profile Technique”, IEEE, 2016, pp. 196-201.

[4] Narasimha Reddy Soora and Parag S. Deshpande, “A novel local skew correction and segmentation approach for printed multilingual Indian documents”, Elsevier, 2018, pp. 1609-1618.