A Novel Approach for Word Segmentation in Correlation based OCR System

Sonam Jain
M-Tech. (CSE)
LLRIET,Moga

Harwinder Singh Sohal
Assistant Professor, IT Deptt
LLRIET,Moga

ABSTRACT

This paper introduces a novel approach for word segmentation in OCR system. Segmentation is one of the substantial sub-processes of the OCR system. The meaning of the word can be changed if segmented word is not correct. An approach of segmentation is formulated in which textual area of image is crimped as one large window .Then large window is divided into small windows of different lines and words are segmented out of each line as sub windows to each small window. Then characters are segmented from sub-windows for recognition. The proposed word segmentation technique works efficiently for variable word spaces.

Keywords

Word Segmentation, OCR, Recognition.

1. INTRODUCTION

In recent years Optical Character Recognition is major field of research in pattern recognition. Optical Character Recognition is famous for processing scanned printed document.OCR is the process of converting text of the image into a machine editable format [3]. The scanned image is altered only after converting it into text. This forms the bases of OCR theory. [6] OCR system model is mainly divided into three sub processes.1) Preprocessing of scanned image. 2) Segmentation text of preprocessed image. 3) Recognition or verifying the text of the image. Document scanning is acquiring a digital image of the original document. So in this step, a good quality scanner is preferred to scan the printed documents. Scanned input image is preprocessed by converting input image into gray-scale image. Grey-scale image have color values from 0-256 .Generally in OCR, optical scanners are used. Printed documents have black print on a white background.

In preprocessing multilevel colored image is converted into bi-level binary image and then noise is eliminated from the binary image. In Segmentation process the size of the character in the text of the preprocessed image is cropped to that of size of character in the template. Recognizing process includes very complex algorithms and uses preloaded characters in database to crosscheck the characters in the segmented text .Verifying process is done either randomly or temporal by human actions. In OCR technology performance of the OCR is highly depends on the quality of the input image. The quality of input image is depends up on the quality of the scanner. Scanner with good colors quality and with high speed is preferred. [4]. Segmentation is the important step in the OCR system because it identifies the characters from a given binary input image [10].Segmentation based the rapidly growing computational power enables the implementation of the present CR methodologies and creates an increasing demand on many emerging application domains, systems are divided into two type's dissection based and recognition-based. The basic principle of dissection is to divide the image into sequence of sub images. Every sub-image is treated as a single character for purpose of recognition. Some of the common dissection techniques used by OCR systems are projection analysis, connected component processing, white space and pitch finding. Dissection technique is suitable for scripts that have spaces between the characters. Recognition-based approach is also known as segmentation – free technique. In this technique it uses a mobile window of variable width that provides the tentative segmentations which are crosschecked by the preloaded database. [3]

2. RELATED WORK

According to a survey of vast literature done by **Casey et.al (1999) and according to Rajiv Kumar et.al (2010) [1]** there are three elementary strategies for segmentation that are as follows:

The Classical Approach, in this segmentation is identified based on characters properties. The process of crimping the image into meaningful components is called dissection.

Recognition Based Segmentation, in this the system searches the image for components that match characters in the preloaded database.

Holistic Methods, in which the system recognizes the words as a whole, thus segmentation into characters is avoided.

A. Cheung. et al. (2001) [4] described that Optical Character Recognition (OCR) systems improve human machine interaction and are widely used in many areas. The recognition of cursive scripts is a difficult task as their segmentation suffers from serious problems. They proposed an Arabic OCR system, which uses a recognition-based segmentation technique to overcome the classical segmentation problems. A newly developed Arabic word segmentation algorithm is also introduced to separate horizontally overlapping Arabic words/sub-words. There is also a feedback loop to control the combination of character fragments for recognition. The system was implemented and the results show 90% recognition accuracy with a 20 chars/s recognition rate.

Nafiz Arica and Fatos T. Yarman-Vural (2001) [12] presented Character Recognition (CR) has been extensively studied in the last half century and progressed to a level sufficient to produce technology driven applications. Now, which require more advanced methodologies. This material serves as a guide and update for readers working in the CR area. First, the historical evolution of CR systems is presented.

Then, the available CR techniques with their superiorities and weaknesses are reviewed. Finally, the current status of CR is discussed and directions for future research are suggested. Special attention is given to the off-line handwriting recognition since this area requires more research to reach the ultimate goal of machine simulation of human reading.

G S Lehal and Chandan Singh (2002) [13] introduced that the post-processing system for OCR of Gurumukhi script has been developed. Statistical information of Punjabi language syllable combinations, corpora look-up and certain heuristics based on Punjabi grammar rules have been combined to design the post-processor. An improvement of 3% in recognition rate from 94.35% to 97.34% has been reported on clean images using the post-processing techniques.

Rajiv Kumar and Amardeep Singh (2010) [6] described that segmentation process is the back bone of the overall OCR process. They said that the segmentation process is the most significant process because if the segmentation is incorrect then they cannot have the correct results; it is just like garbage in and garbage out. But it is not an easy job, because segmentation is one of the complex processes. It is more difficult if the document is handwritten because in that case only few points are there which can be used to make segmentation. They formulated an approach to segment the scanned document image. Initially they considered the whole image as one large window. Then this large window is broken into less large windows giving lines, once the lines are identified then each window consisting of a line is used to find a word present in that line and finally to the characters.

Nikos Nikolaou et al. (2010) [7] described segmentation technique that works efficiently for documents with non-constant spaces between text lines, words and characters. The word segmentation technique is based on the construction of histograms between the horizontal distances between adjacent bounding boxes. The adjacent connected components with distance smaller than the threshold are considered to belong to the same word. Threshold value is calculated to the maximum value that represents the peak of the histogram.

Safwa Taha et al. (2012) [8] introduced the segmentation algorithm based on three levels of segmentation: lines segmentation, sub-words segmentation and characters segmentation. In line segmentation lines are segmented by creating horizontal projection profiles of the image rows to find the empty rows between the rows containing the text. Words segmentation is the next level after the lines are segmented. Vertical projection profile of images columns are used to divide lines into words. Character segmentation is the third level of segmentation which is based on features of Arabic cursive text.

3. PRESENT WORK

Input scanned image is processed using OCR system model. In this system model image is preprocessed by using binarization and noise elimination. After that segmentation is done which consists of line segmentation, character segmentation following by new technique of word segmentation. At last Recognition is done to correlate the scanned image with the database and find the template with maximum correlate value to recognize it as character of the scanned document .The recognized characters are written into a text file.

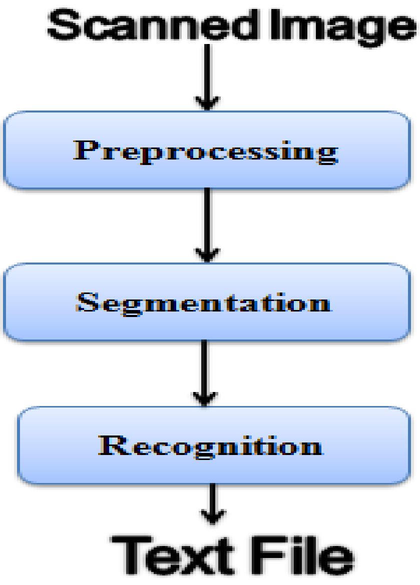


Figure 1: The System Model

3.1 Preprocessing

Image is preprocessed before segmentation. It is done by image binarization and noise elimination from the input scanned image. Image binarization is to convert the input image into binary image.

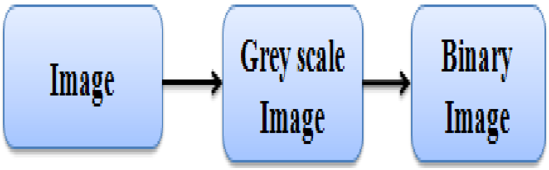
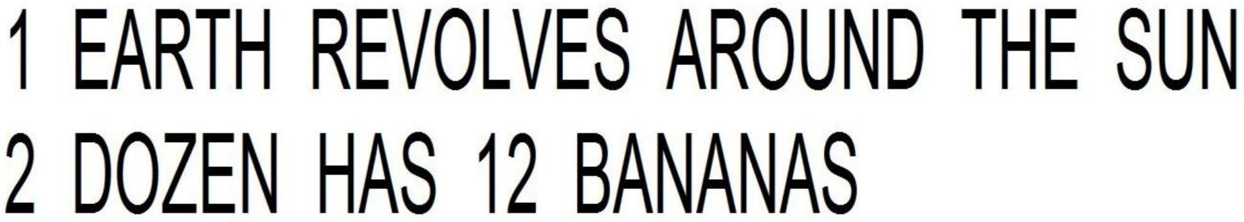


Figure 2: Image Binarization

Noise elimination is the smoothing of the image by applying filters, which will increases performance of OCR. Hence, more the image is noise free more the accuracy of OCR system. In this work filters are applied on the binary image because in that case color values are either 0 or 1. Through rather than the abrupt change in case of colored image there will be a gradual change which leads to smoothing instead of blurring of the image.

3.2 Segmentation

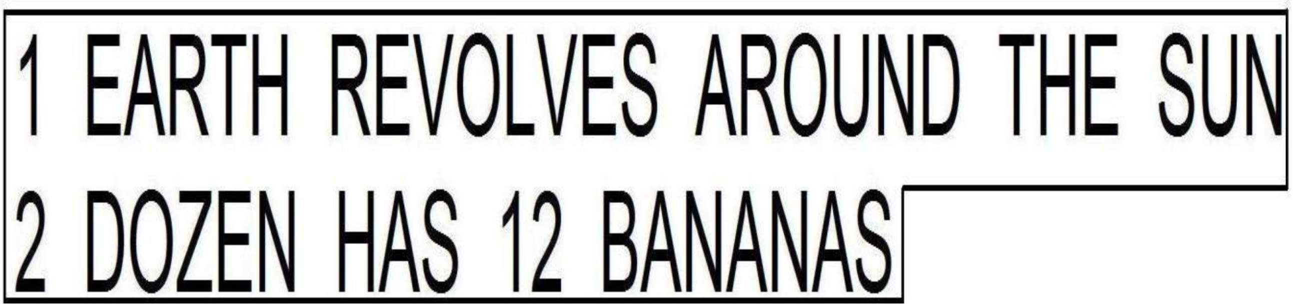
In English language there are 26 characters and template of each character is created as database, which is used during segmentation and recognition process. During segmentation preprocessed image is crimped according to the textual area of the image and it is the one large window. Following figure shows the example of the Preprocessed Scanned Image.



1 EARTH REVOLVES AROUND THE SUN
2 DOZEN HAS 12 BANANAS

Figure 3: Preprocessed Scanned Image

These two lines are crimped according to textual area. Image is crimped by using horizontal projection means pixels which are having 1 pixel values are only considered as start point and end point of the crimped image. In this case image is crimped from number 1 of the first line to alphabet S of the last line as shown in the following figure.

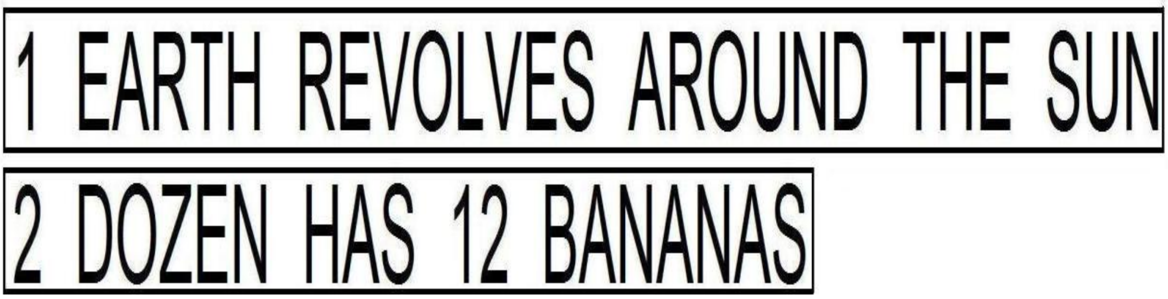


1 EARTH REVOLVES AROUND THE SUN
2 DOZEN HAS 12 BANANAS

Figure 4: Crimped Image

3.2.1 *Line segmentation*

The lines are segmented by crimping first line from number 1 to alphabet N. First line is determined by using values of the pixels .In horizontal projection to the crimped image after alphabet N at next row pixel value is zero and if pixel value is zero then first line is crimped up to when there is pixel value is 1.



1 EARTH REVOLVES AROUND THE SUN
2 DOZEN HAS 12 BANANAS

Figure 5: Line Segmentation

3.2.2 *Word Segmentation*

In this approach of word segmentation the distance between all the characters of line is calculated and then a threshold value is set for comparing each character space. If the space of the character is less than the threshold value then this is not a word space and it is the character space. If word space varies while typing then for this word space is set in range of 4 less than the calculated word space to the maximum calculated distance between two. This algorithm works more efficiently when there is variable space between different words of English language. The variability between spaces is due to shape of different characters of English language. For Example shape of character “I” is different in pixel value to the shape of character “M”. Hence, this approach of word segmentation is more efficient.

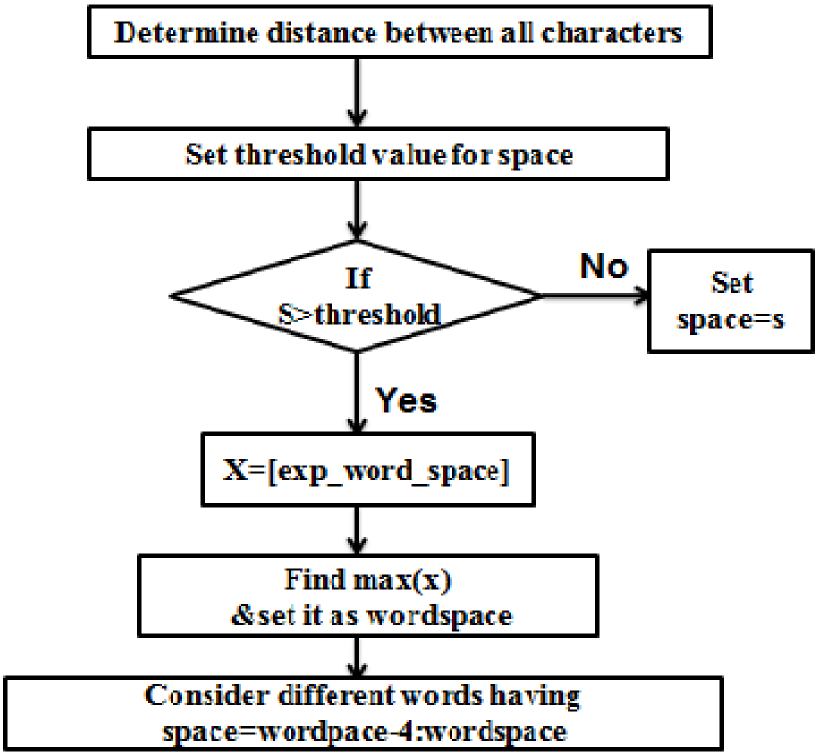

Determine distance between all characters
Set threshold value for space
If S>threshold
No
Set space=s
Yes
X=[exp_word_space]
Find max(x) &set it as wordspace
Consider different words having space=wordpace-4:wordspace

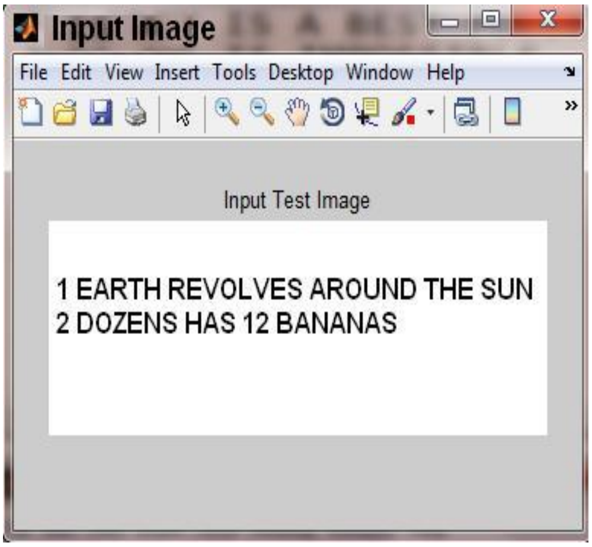
Figure 6: Flow Chart of Word Segmentation

3.2.3 Character Segmentation

In character segmentation the position of the object i.e., the character in the image is found out and the size of the image is cropped to that of the template size. The segmented image is correlated with the template, which are preloaded into the system. Once the correlation is completed, then the template with the maximum correlated value is declared as the characters present in the scanned image

4. RESULTS

4.1 Input Image

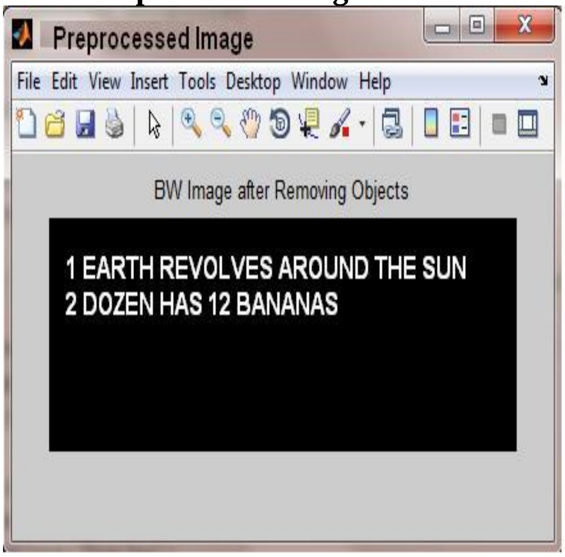


Input Test Image

1 EARTH REVOLVES AROUND THE SUN
2 DOZENS HAS 12 BANANAS

Figure 7: Input Image

4.2 Preprocessed Image



BW Image after Removing Objects

1 EARTH REVOLVES AROUND THE SUN
2 DOZEN HAS 12 BANANAS

Figure 8: Preprocessed Image

4.3 Segmented Images

The text of the image is segmented into different sub-windows as shown in Figures below

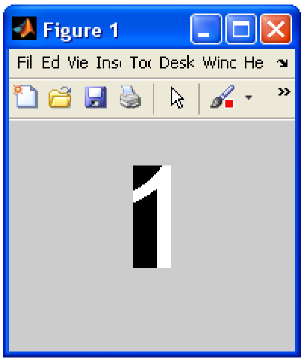


Figure 9: First Word Segmentation of first line

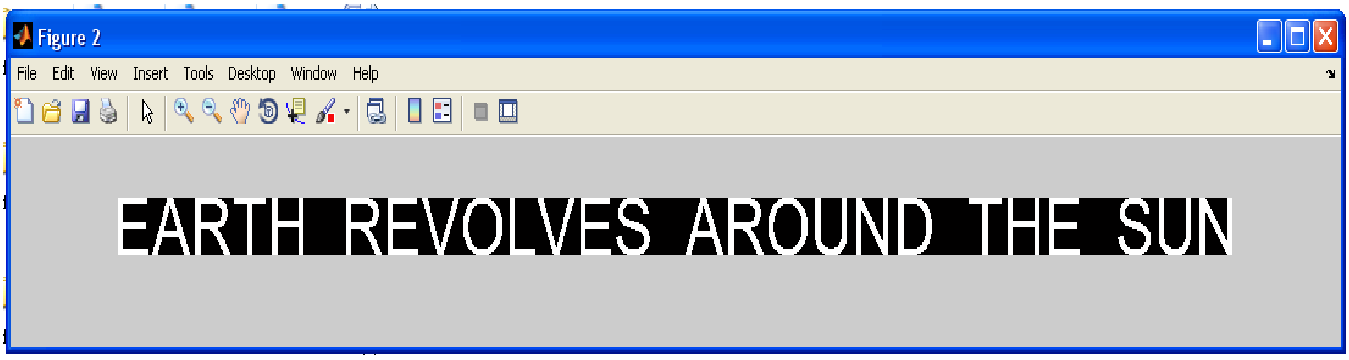


Figure 2
File Edit View Insert Tools Desktop Window Help
EARTH REVOLVES AROUND THE SUN

Figure 10: Remaining Word Segmentation of first line

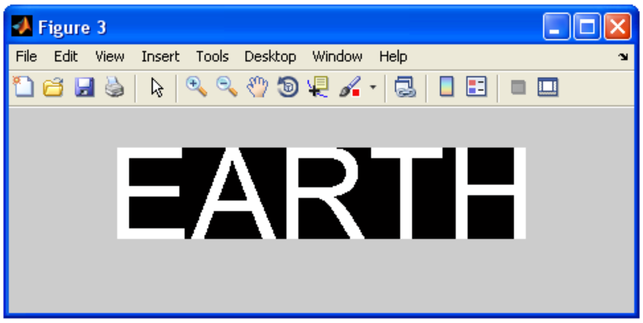


Figure 11: Second Word Segmentation of first line

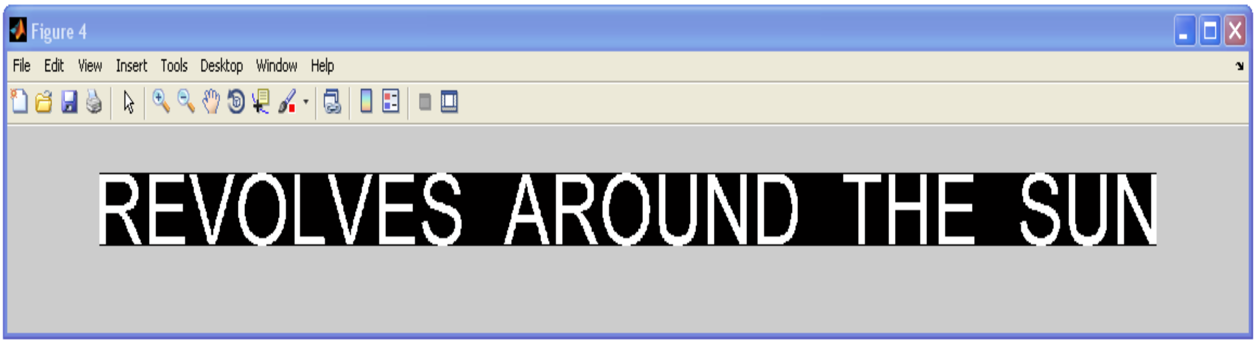


Figure 12: Remaining Word Segmentation of first line

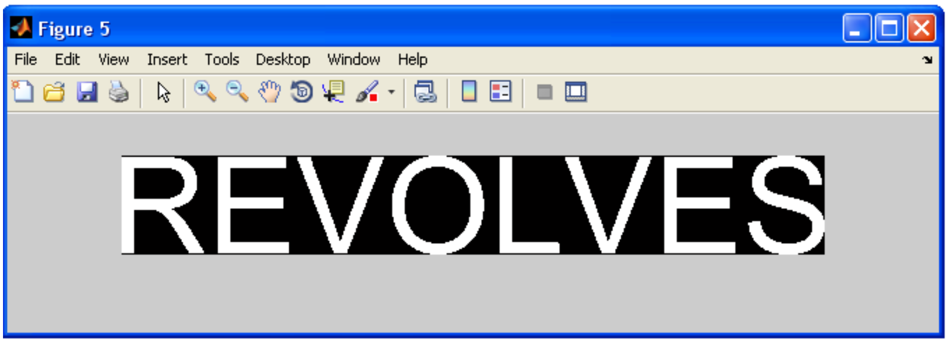


Figure 13: Third Word Segmentation of first line

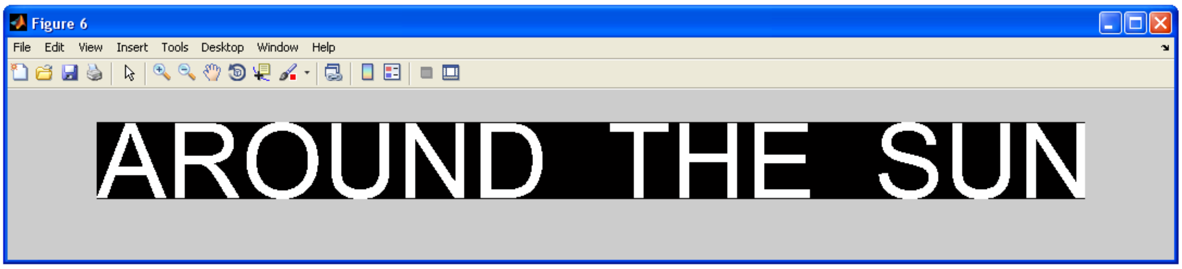


Figure 14: Remaining Word Segmentation of first line

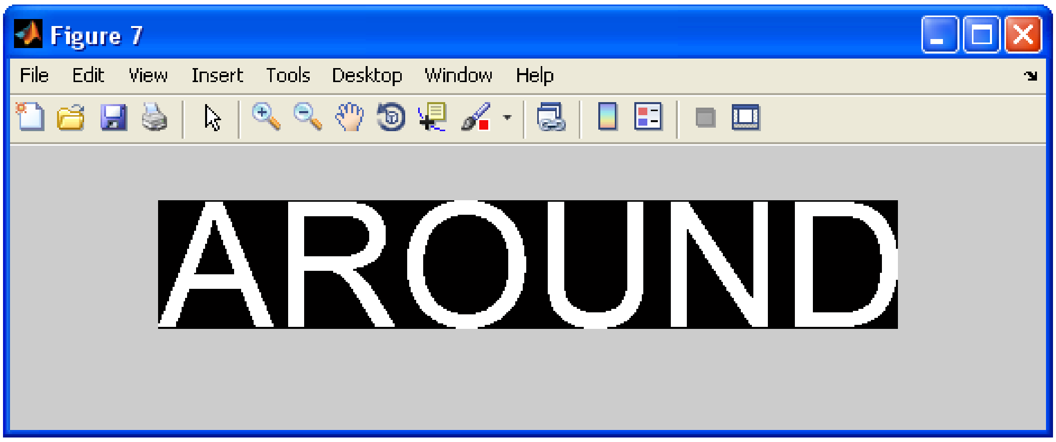


Figure 15: Fourth Word Segmentation of first line

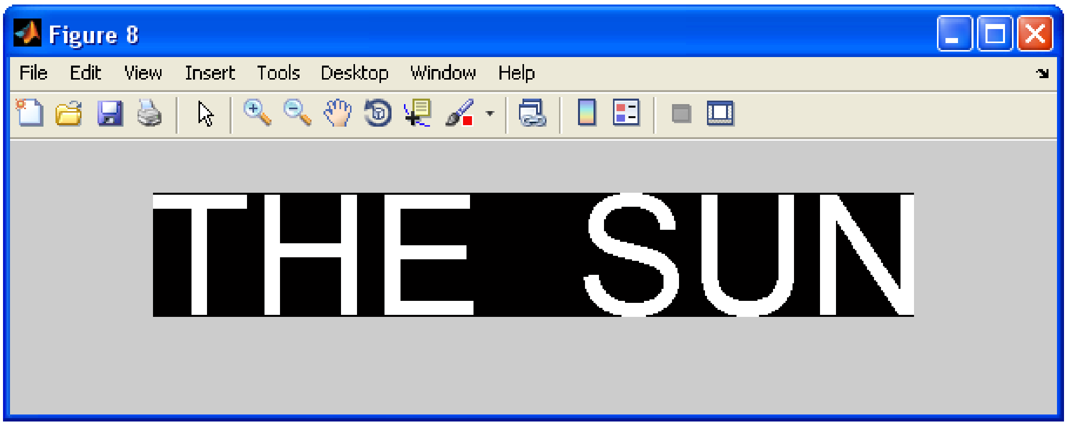


Figure 16: Remaining Word Segmentation of first line

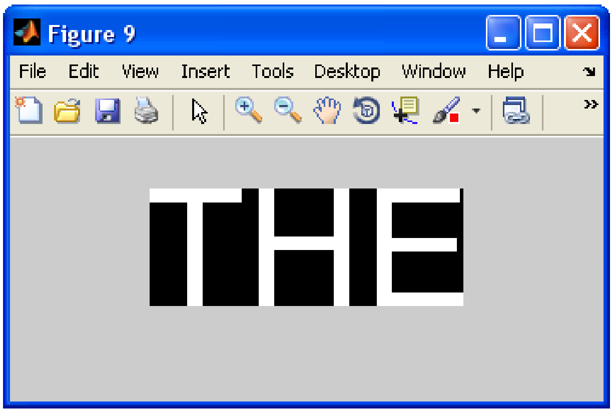


Figure 17: Fifth Word Segmentation of first line

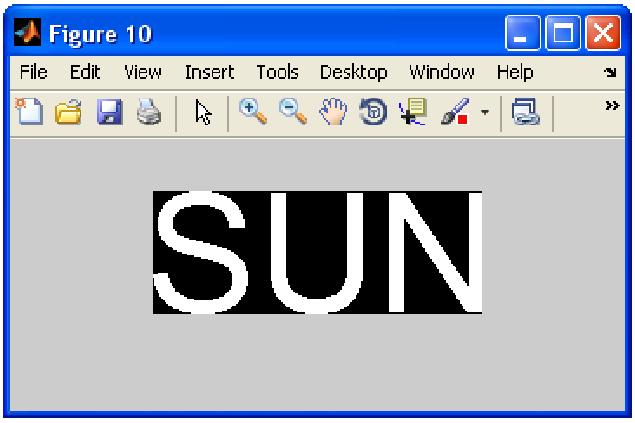


Figure 18: Remaining & Last Word Segmentation of first line

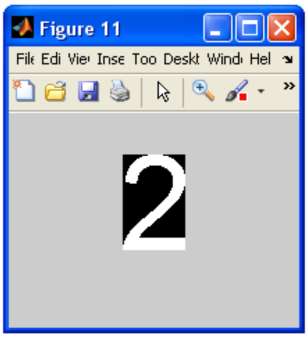


Figure 19: First Word Segmentation of Next line

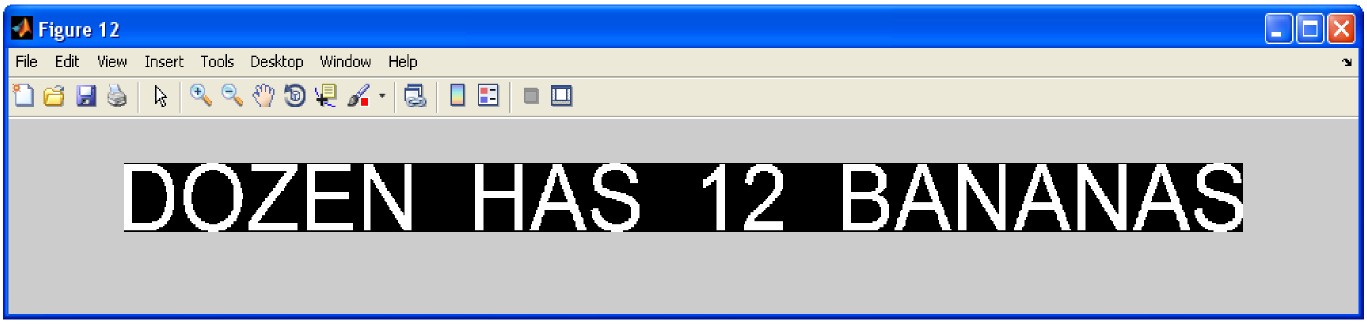


Figure 20: Remaining Word Segmentation of Next line

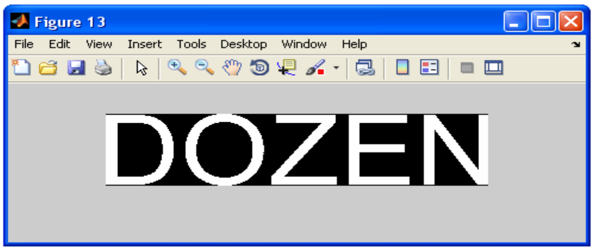


Figure 21: Second Word Segmentation of Next line

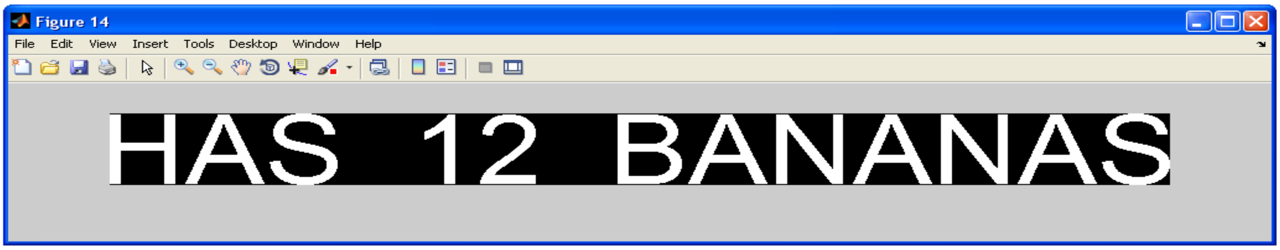


Figure 22: Remaining Word Segmentation of Next line

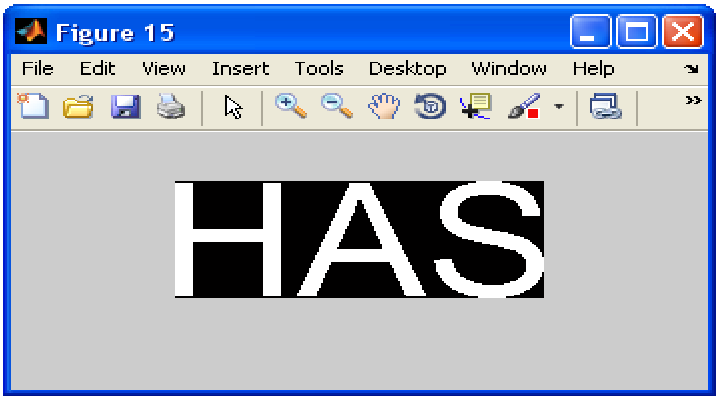


Figure 23: Third Word Segmentation of Next line

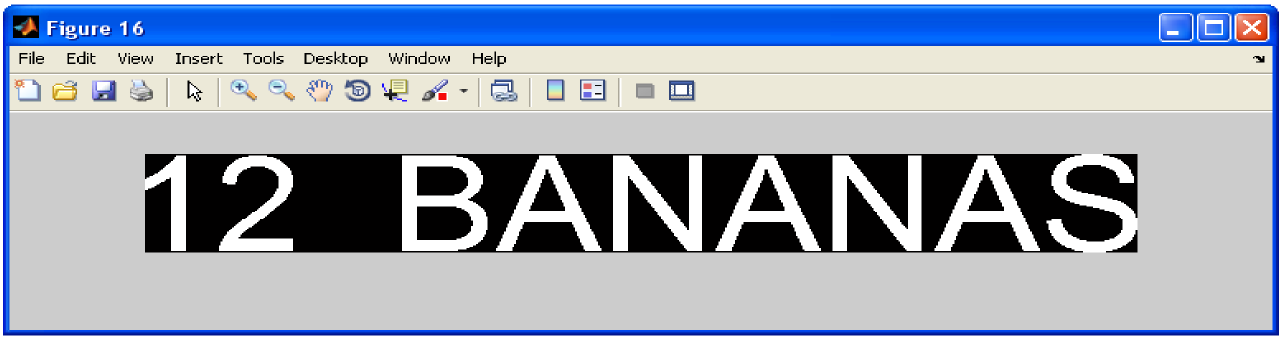


Figure 24: Remaining Word Segmentation of Next line

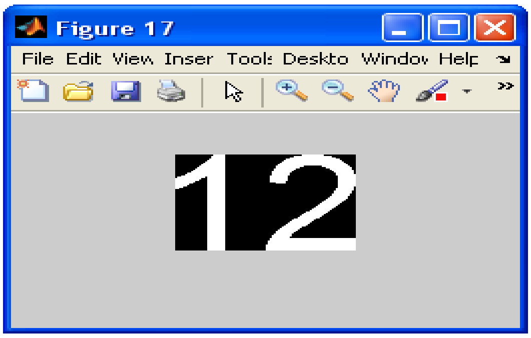


Figure 25: Fourth Word Segmentation of Next line

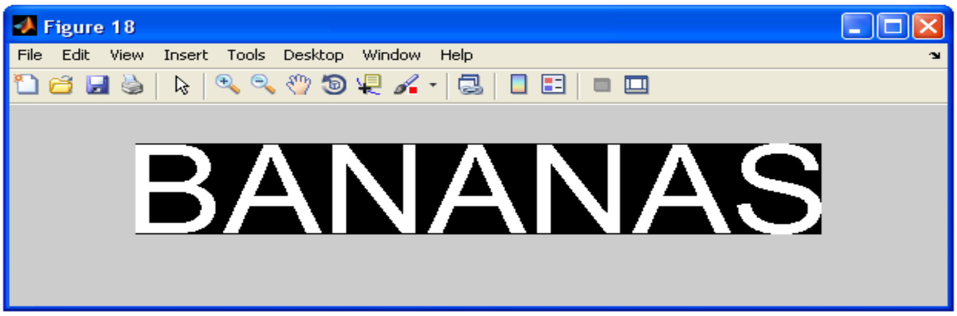


Figure 26: Remaining & Last Word Segmentation of Next line

5. CONCLUSION AND FUTURE WORK

After collecting the results this technique is implemented on various scanned text images. Words are segmented to accuracy of 100% with this new technique of word segmentation and are recognized with efficiency of 99%. This technique is used for English language in which characters are having distinguish space between them .Future scope of this work is to implement this technique on different Indian scripts. India is a multi-language country where almost 12 different scripts are used. Future of this is to implement this method on Gurumukhi Script

6. REFRENCES

[1]. Casey, R.G. and Lecolinet, E., "A Survey of Methods and Strategies in Character Segmentation**",** IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.18, no.8, pp.690-706, 1996.

[2]. Issam Bazzi, Richard Schwartz and John Makhoul "An Omnifont Open-Vocabulary OCR Systemfor English and Arabic" IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 21, No. 6, 1999.

[3]. Rejean Plamondon, Sargur N. Srihari "On-Line and Off-Line Handwriting Recognition:A Comprehensive Survey" 1EEE Transactions On Pattern Analysis And Machine Intelligence. Vol. 22 , No. 1,2000.

[4]. A Cheung, M. Bennamoun, N.W. Bergmann "An Arabic Optical Character Recognition system using recognition-based segmentation" Pattern Recognition Society. Published by Elsevier Science Ltd., pp 215-233, 2001.

[5]. N. Arica and Fatos T. Yarman-Vural "An Overview of Character Recognition Focused on Off-Line Handwriting" IEEE Transactions On Systems, Man And Cybernetics—Part C: Applications And Reviews, Vol. 31, No. 2,2001.

[6]. Rajiv Kumar, Amardeep Singh "Detection and Segmentation of Lines and Words in Gurmukhi Handwritten Text" IEEE, 2nd International Advance Computing Conference, pp 353-356, 2010.

[7]. Nikos Nikolaou,Michael Makridis,Basilis Gatos, Nikolaos Stamatopoulos , NikosPapamarkos "Segmetat-ion of historical machine-printed documents using Adaptive Run Length Smoothing and skeleton segmentation paths"ELSEVIER, Image and Vision Computing 28 2010) 590–604.

[8]. Pranob K Charles, V.Harish, M.Swathi, CH. Deepthi "A Review on the Various Techniques used for Optical Character Recognition" International Journal of Engineering Research and Applications (IJERA) , Vol. 2, Issue 1 , pp.659-662 659, 2012.

[9]. Safwa Taha, Yusra Babiker and Mohamed Abbas "Optical Character Recognition of Arabic Printed Text" IEEE Student Conference On Research And Development,pp 235-240,2012

[10]. Gaurav Singla, Dr. Parmod Kumar "Extract the Punjabi Word from Machine Printed Document Images" Int. Journal of Engineering Research and Application Vol. 3, Issue 5, pp.343-348,2013.

[11]. Ravina Mithe, Supriya Indalkar, Nilam Divekar **"**Optical Character Recognition" International Journal of RecentTechnology and Engineering (IJRTE), Vol-2, Issue-1, 2013.

[12]. Nafiz Arica and Fatos T. Yarman-Vura (2001) "An Overview of Character Recognition Focused on Off-Line Handwriting" IEEE Transactions On Systems, Man And Cybernetics—Part C: Applications And Reviews, Vol. 31, No. 2.

[13]. G S Lehal and Chandan Singh (2002), "A post-processor for Gurmukhi OCR" Sadhana, vol. 27, pp. 99–111.