International Journal of Science Engineering and Advance Technology

# Recognition of characters in document images using morphological operation

**P. EswaraBabu, K. Ravi Kumar, Dr. G. Lavanya Devi**
Final M.Tech Student, Assoc. Professor
Dept. of Computer Science and Engineering,
Kakinada Institute of Engineering & Technology, Korangi, E.g. A.P.
Asst. prof, Dept. of CSS college of Engineering, Andhra University, Visakhapatnam, A.P.

***Abstract*** - In this paper, we deal with the problem of document image rectification from image captured by digital cameras. The improvement on the resolution of digital camera sensors has brought more and more applications for non-contact text capture. It is widely used as a form of data entry from some sort of original paper data source, documents, sales receipts or any number of printed records. It is crucial to the computerization of printed texts so that they can be electronically searched, stored more compactly, displayed on-line, and used in machine processes such as machine translation, text-to-speech and text mining. Unfortunately, perspective distortion in the resulting image makes it hard to properly identify the contents of the captured text using traditional optical character recognition (OCR) systems. In this work we propose a new technique; it is a system that provides a full alphanumeric recognition of printed or handwritten characters at electronic speed by simply scanning the form. Optical character recognition, usually abbreviated as OCR is the mechanical or electronic conversion of scanned images of handwritten, typewritten or printed text into machine-encoded text. OCR software detects and extracts each character in the text of a scanned image, and using the ASCII code set, which is the American Standard Code for Information Interchange, converts it into a computer recognizable character. Once each character has been converted, the whole document is saved as an editable text document with a highest accuracy rate of 99.5 per cent, although it is not always this accurate. The basic idea of Optical Character Recognition (OCR) is to classify optical patterns (often contained in a digital image) corresponding to alphanumeric or other characters.

***Keywords:*** Document image analysis; Document image rectification; Optical character recognition; Morphological image processing; ASCII code set.

## I. INTRODUCTION

Since late 1920s, there have been attempts by many engineers to develop OCR systems .The engineering attempts at automated recognition of printed characters started prior to World War II. However, it was not until the 1950s that the first commercial OCR system became available. This was because the technology was not needed in many places, and it was too expensive to implement OCR due to its immature algorithms during the era.

Document scanner is widely used to capture text and transform it into electric form for further processing. Ascamera resolution rises in recent years, high-speed noncontact text capture through digital cameras is becoming an alternative choice. Unfortunately, perspective distortion coupled with captured images by digital camera brings up a new problem to the traditional optical character recognition (OCR) system. Similar to the skew compensation operation required after the scanning process, perspective distortion must be removed before the document image is fed to the OCR system.

A proliferation of research has been dedicated to the detection and correction of document skew that is introduced during the scanning process. The proposed correction approaches can be roughly classified into three categories, namely, Hough transform based method [1, 2], projection profile based method [3], and nearest neighbour based method [4]. One common feature of these methods is that they all assume that skew distortion is rotation-induced, which means that the top line and base line of each text line are still parallel to each other within the scanned document image. Therefore, none of them can handle the perspectively distorted documentimages where the parallel relation between top line and base line of text lines is totally destroyed.

Optical Character Recognition (OCR) is a type of document image analysis where scanned digital

image that contains either machine printed or handwritten script input into an OCR software engine and translating it into an editable machine readable digital text format (like ASCII text).OCR works by first pre-processing the digital page image into its smallest component parts with layout analysis to find text blocks, sentence/line blocks, word blocks character blocks. Other features such as lines, graphics, photographs etc are recognized and discarded. The character blocks are then further broken down into components parts, pattern recognized and compared to the OCR engines large dictionary of characters from various fonts and languages. Once a likely match made then this is recorded and a set of characters in the word block are recognized until all likely characters have been found for the word block. The word is then compared to the OCR engine's large dictionary of complete words that exist for language.

The segmentation algorithm, proposed in this study, segments the whole word into strokes, each of which corresponds mostly to a character or rarely to a portion of a character. Recognition of each segment is accomplished in three stages: In the first stage, characters are labelled in three classes as ascending, descending, and normal characters. In the second stage, Hidden Markov Model (HMM) is employed for shape recognition. The features extracted from the strokes of each segment are fed to a left-right HMM. The parameters of the feature space are also estimated in the training stage of HMM. Finally, an efficient word-level recognition algorithm resolves handwriting strings by combining lexicon information and the HMM probabilities.

## II. SYSTEM OVERVIEW

This paper provides a general survey and basic implementation of Optical Character Recognition. First, the history and current state of OCR technology is examined. Then, there is an overview of the methods employed by OCR programs and the classification algorithms therein. Finally, there is a basic MATLAB implementation of an OCR program that will take text in an image and convert it to plain text.

### 2.1 Phases Involved

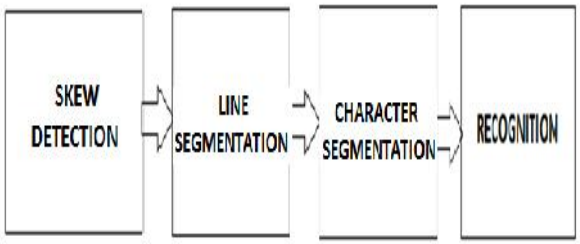



**Fig.2.1 Phases involved in the project**

Prior to the skewing process the image is first preprocessed to make it ready for skew detection, As the captured image is colored in nature, it is required to convert it into a gray image with intensity levels varying from 0 to 255 (8-bit image). Then it is converted into a binary image with suitable threshold (Black=0 & White=1). The advantage is that the handling of the image for further processing becomes easier. This binary image is then inverted i.e. black is made white and white is made black. By doing so, the segmentation process becomes easier.

#### 2.1.1 Skew Detection

OCR systems mainly depend upon the accuracy of pre-processing stage. The digital image of a document may be skewed/ rotated arbitrarily because of how it was placed on the platen when it was scanned or because of a document feeder malfunction. A significant skew in document can be detected by human vision easily and the skew correction can be made by re-scanning the document, whereas for mild skew it may not be possible to notice its skew as human vision system fails to identify it. Even a smallest skew angle existing in a given document image results in the failure of segmentation of complete characters from words or a text lines, as the distance between the character reduces.

#### 2.1.2 Line Segmentation

When the image matrix is ready to be processed, the first step is to isolate each line of the text from the whole document. A horizontal projection profile technique is used for this Purpose. A computer program scans the image horizontally to find the first and last black pixels in a line. Once these pixels are found, the area in between these pixels represents the line that may contain one or more character. Using the same technique, the whole document is scanned and each line is detected and saved in a temporary array for further processing.

#### 2.1.3 Character Segmentation

Once each line of the text is stored in a separate array, using vertical projection profile, the program scans each array this time vertically to detect and isolate each character within each line. The first and last black pixels that are detected vertically are the borders of the character. It possible that when the characters are segmented, there is a white area above, below, or both above and below the character, except for the tallest character that its height is equal to the height of the line.

#### 2.1.4 Recognition

In the recognition process, each character extracted is correlated with each and every other character present in the database. The database is a predefined set of characters. All the characters in

the database are resized to 42x24. By knowing the maximum correlated value, from the database, the character is identified. Finally, the recognized characters are made to display on a notepad.**2.2**

**Description of OCR Block Diagram**

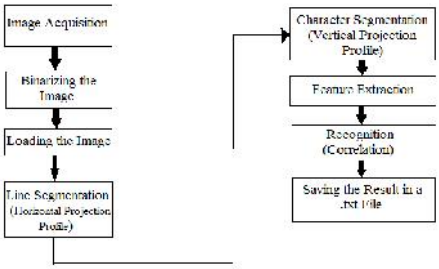



**Fig.2.2 Block Diagram of OCR**

**2.3 Flowchart of OCR**

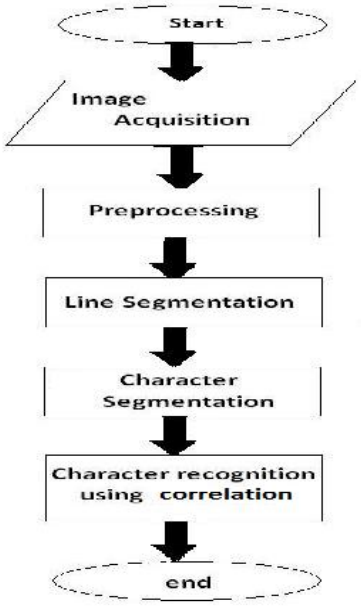



**Fig.2.3 Flowchart for OCR**

## III. SKEW DETECTION AND SEGMENTATION

In this section, it is illustrated about skew detection and segmentation of the scanned image. Intensive research work in the field of skew detection has given birth to many methods. The basic steps involved in Skew Detection using Hough Transform are

i) Reading of an Scanned image
ii) RGB to Gray Conversion
iii) Gray to Binary Conversion
iv) Image negative
v) Morphological techniques
vi) Skew Compensation

### 3.1 Reading of an Scanned Image

It reads a gray scale or colour image from the file specified by the string filename. If the file is not in the current directory, or in a directory on the MATLAB path, specify the full pathname.

#### 3.1.1 Format of an Image

Various scanning packages can handle different formats. The various formats of an image are BMP — Windows Bitmap
JPEG — Joint Photographic Experts Group
PNG — Portable Network Graphics
TIFF — Tagged Image File Format
GIF — Graphics Interchange Format
PBM — Portable Bitmap


**Fig.3.1 JPEG Image Format**

In general, as we are working on Microsoft Windows System we use BMP or JPEG or TIFF files. By default, Microsoft Windows cursors are 32-by-32 pixels. MATLAB pointers must be 16-by-16.So we will probably need to scale our image by using imresize function.

### 3.2 Conversion of Image

After reading a scanned image which is either in JPEG or BMP format, it should undergo the following three conversion process before applying the Morphological Techniques and they are

1) RGB to GRAY Conversion
2) GRAY to BINARY Conversion
3) Image Negative


**Fig.3.2 Gray Scale Image**


**Fig.3.3 Binary Image Format**


**Fig.3.4 Image Negative Format**

### 3.3 Morphological Techniques

A related problem facing character recognition systems is the separation of text from a textured or decorated background. The most common basic morphological operations used in image processing are *dilation and erosion*.

In dilation, if at least one of the pixels in the structuring element is set to a foreground value and its corresponding image pixel is also set to a foreground value, then the centre of the image is set to foreground (note that a separate lighter colour is used to indicate new foreground pixels, but in an actual implementation, each of these pixels would be assigned the same colour intensity).

In erosion, the centre of the image is only set to foreground if at least one structuring element is set to foreground and all of the structuring element pixel values exactly match their corresponding image location pixels. Dilation tends to have the effect of increasing the foreground area of an image, often filling in or smoothing small holes. Erosion on the other hand tends to have the opposite effect; decreasing the foreground area of the image, and increasing the size of holes. Combining these operations by performing erosion, followed by a dilation on the eroded image is called an opening operation, and can often be used to eliminate small textured backgrounds from foreground text to reasonably good effect.

a) In a morphological operation, the value of each pixel in the output image is based on a comparison of the corresponding pixel in the input image with its neighbors.
b) **Dilation:** Value of the output pixel is the maximum value of all the pixels in the input pixel's neighborhood. In a binary image, if any of the pixels is set to the value 1, the output pixel is set to 1.
c) **Erosion:** Value of the output pixel is the minimum value of all the pixels in the input pixel's neighborhood. In a binary image, if any of the pixels is set to 0, the output pixel is set to 0.
d) **Close:** Performs morphological closing i.e., dilation followed by erosion.
e) **Thicken:** With n = Inf, thickens objects by adding pixels to the exterior of objects until doing so would result in previously unconnected objects being 8-connected.

### 3.4 Skew Compensation by using Hough Transform

A common problem that occurs when a flatbed scanner is employed to digitize paper documents is that the paper is often placed so that it does not lie exactly perpendicular with the scanner head. Instead it is often rotated some arbitrary angle, so that when scanned the resultant digitized document image appears skewed. Depending on the algorithms employed, working with skewed documents directly can lead to difficulties when attempts are made to segment the input image into columns, lines, words or individual character regions. One simple illustration of this is the (naive) attempt of creating a line finder that works by summing pixel intensity values horizontally across each row of pixels, then scanning these sums vertically to find consecutive valleys or runs of minimal sum (this assumes the document to be recognized has a horizontal reading order).

**3.5 Segmentation**

Once a page has been suitably digitized, cleaned up, and had its textual regions located, it is ready to be segmented so that the individual symbols of interest can be extracted and subsequently recognized.

Segmentation refers to a process of partitioning an image into groups of pixels which are homogeneous with respect to some criterion. Segmentation algorithms are area oriented instead of pixel oriented. The result of segmentation is the splitting up of the image into connected areas. Thus segmentation is concerned with dividing an image into meaningful regions. Image segmentation can be broadly classified into two types [3]

i) Local Segmentation: It deals with the segmenting sub images which are small windows on a whole image.
ii) Global segmentation: It deals with the images consisting of relatively large number of pixels and makes estimated parameter values for global segments more robust.

In this chapter for character segmentation, first the image has to be segmented row-wise (line segmentation), then each rows have to be segmented column-wise (character segmentation). Finally, characters can be extracted using suitable algorithms such as edge detection technique; histogram based methods or connected component analysis method. We also describe our approach to grouping these isolated images together so that there is only one (or a few) such representatives for each character symbol. It is at this stage, that our character recognition strategy begins to distinguish itself from classical or historically driven approaches to recognition. Such systems typically do not make attempts to cluster similar shapes together, instead they move directly to the recognition of isolated character images (via image shape features).

In our project segmentation process is divided into two parts. They are

a) Line Segmentation
b) Character Segmentation

**3.5.1 Line Segmentation**

When the image matrix is ready to be processed, the first step is to isolate each line of the text from the whole document. A Horizontal Projection Profile technique is used for this purpose. The line segmentation is carried out by scanning the entire row one after the other and taking its sum. Since black is represented by 0 and white by 1, if there is any character present, then the sum would be non-zero. Thus the line segmenting is carried out. Once these pixels are found, the area in between these pixels represents the line that may contain one or more characters. Using the same technique, the whole document is scanned and each line is detected and saved in a temporary array for further processing.

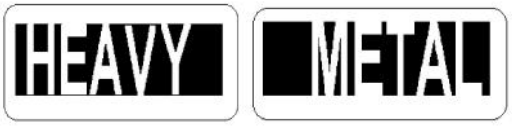



**Fig.3.6 Line segmented image of printed text**

**3.5.2 Character Segmentation**

Once each line of the text is stored in a separate array, using Vertical Projection Profile, the program scans each array this time vertically to detect and isolate each character within each line. The first and last white pixels that are detected vertically are the borders of the character.

If in one vertical scan two or less black pixels are encountered then the scan is denoted by 0, else the scan is denoted by the number of white pixels. In this way a vertical projection profile is constructed. Now, if in the profile there exist a run of at least $k1$ consecutive 0's then the midpoint of that run is considered as the boundary of a word.


**Fig.3.7 Characters extracted from a Printed text**

In this chapter it totally provides information about un skewing of the scanned image and segmentation i.e., both line and word segmentation. In next chapter it is explained about character recognition.

## IV. CHARACTER RECOGNITION

In this section, it introduces Character Recognition process. Recognition is the last step in the character recognition process. "Character Recognition" is an offline recognition system developed to identify either printed characters or discrete run-on handwritten characters. It is a part of pattern recognition that usually deals with the realization of the written scripts or printed material into digital form. The main advantage of storing these written texts in digital form is that, it requires less space for storage and can be maintained for further references without referring to the actual script again and again.

Character recognition is a process, which associates a symbolic meaning with objects (letters, symbols and numbers) that are on an image, *i.e.,* character recognition techniques associate a symbolic

identity with the image of a character. Mainly, character recognition machine takes the raw data that further implements the *pre-processing*of any recognition system.

**4.1 Types of Character Recognition**

Character recognition is an extremely large field which can be divided generally into two fields:

a) On-line character recognition
b) Off-line character recognition.

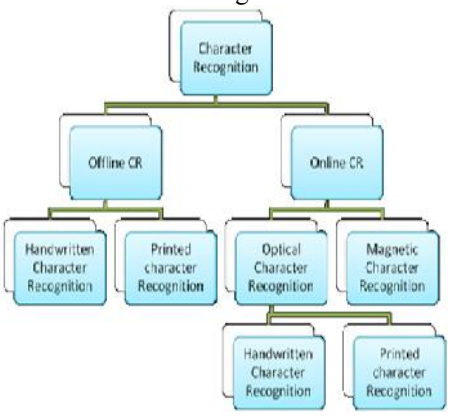



**Fig.4.1: Classification of Character recognition systems**

**4.1.1 On-Line Character Recognition**

On-line recognition is the process of recognizing the characters as they are being written. Also online character recognition has real time contextual information. Examples of systems that employ on-line recognition include the Apple Newton, Palm Pilot, Touch screen mobiles. In case of online handwritten character recognition, the handwriting is captured and stored in digital form via different means. Usually, a special pen (Stylus) is used in conjunction with an electronic surface. As the pen moves across the surface, the two- dimensional coordinates of successive points are represented as a function of time and are stored in order.

**4.1.2 Off-Line Character Recognition**

On the other hand, off-line recognition is a system that recognizes by capturing an image of the characters or handwritten text that are to be recognized. Off-line recognition systems' potential lies in fields such as document processing, mail direction, and cheque verification. Offline handwriting recognition refers to the process of recognizing words that have been scanned from a surface (such as a sheet of paper) and are stored digitally in gray scale format. After being stored, it is conventional to perform further processing to allow superior recognition but offline data does not support for real time contextual information. This difference generates a significant divergence in processing architectures and methods.

The offline character recognition can be further classified into two types:

i) Magnetic Character Recognition (MCR)
ii) Optical Character Recognition (OCR)

i) Magnetic Character Recognition (MCR)

In MCR, the characters are printed with magnetic ink. The reading device can recognize the characters according to the unique magnetic field of each character. MCR is mostly used in banks for cheque authentication service and also for updating entries in the transaction statements

ii) Optical Character Recognition (OCR)

OCR deals with the recognition of characters acquired by optical means, typically a scanner or a camera. The characters are in the form of pixelated images, and can be either printed or handwritten, of any size. OCR can be subdivided into handwritten character recognition and printed character recognition. Handwritten Character Recognition is more difficult to implement than printed character recognition due to diverse human handwriting styles and customs. In printed character recognition, the images to be processed are in the forms of standard fonts like Times New Roman, Arial Black, etc.

## I. EXPERIMENTAL RESULTS

There are different types of techniques used for Optical character recognition. In this chapter results obtained for Optical Character recognition using the correlation technique are discussed**.**

**5.1 Image Acquisition**

The printed test image with skew is shown in the fig 8.1. These images are further processed according to the algorithm.

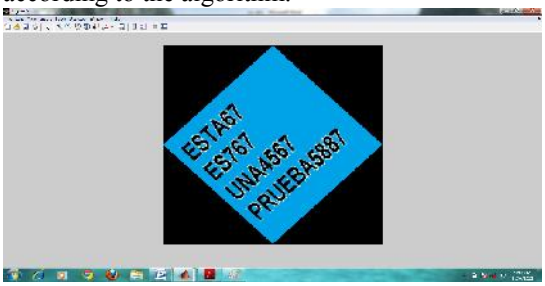



**Fig.5.1 Scanned image of printed text**

**5.1.2 Binarizing the Image**

Binarizing the image includes rgb to gray conversion and then gray to BW image. This can be shown in images shown below:

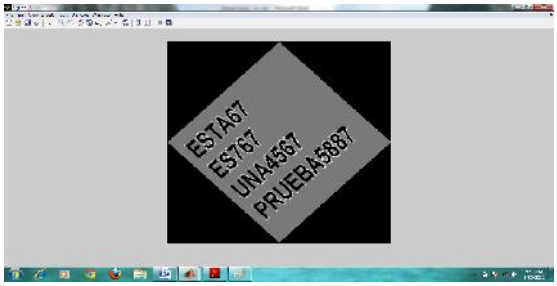



**Fig.5.2 RGB to Gray converted Image**


**Fig.5.3 Black & White Image**

**5.1.3 Morphological Operations**

Morphological operation is applied to binarized image after negating the image i.e., black image with white background. The morphological operations that are applied to the image are 'close' and 'thicken'. The images are as shown below:

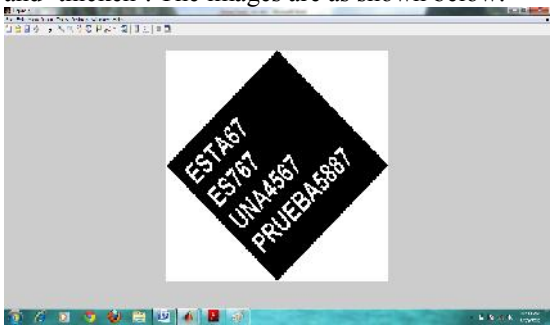



**Fig.5.4 Image negative**


**Fig.5.5 Image after Close operation**

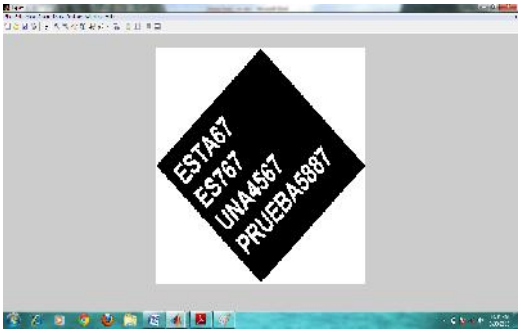



**Fig.5.6 Image after Thicken operation**

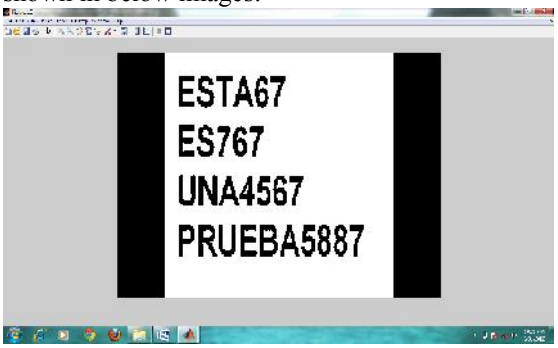
**5.1.4 Deskewing the Image**

After applying morphological techniques the image negative is applied to the image before deskewing. By using Hough transform we will find out the skewing angle of image rotated. By using 'rotate' keyword and obtained angle, original image is deskewed as shown below:


**Fig.5.7 Deskewed image**

**5.2 Results of Cropping an Image**

The deskewed image is cropped row and column wise before applying for segmentation process as shown in below images:



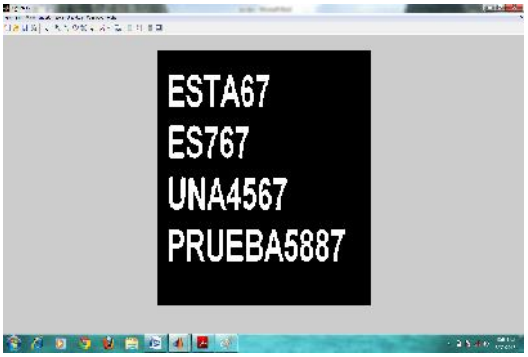




**Fig.5.8 cropping of the deskewing images**

### 5.3 Results of Line Segmentation

The preprocessed images are segmented row-wise (line segmentation). The resulted images of the line segmentation are shown below:

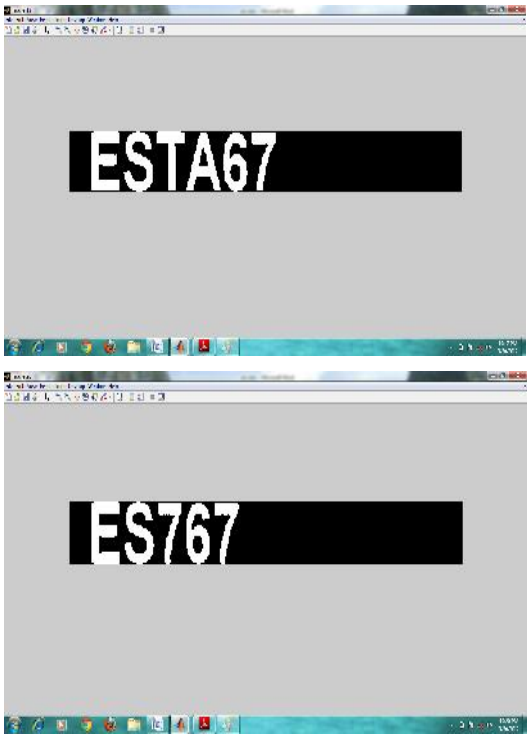



**Fig.5.9 Line segmented images of printed text**

### 5.4 Results of Character Segmentation

From the line segmented image each character is segmented column-wise (character segmentation). The resulted images of the character segmentation are shown below:

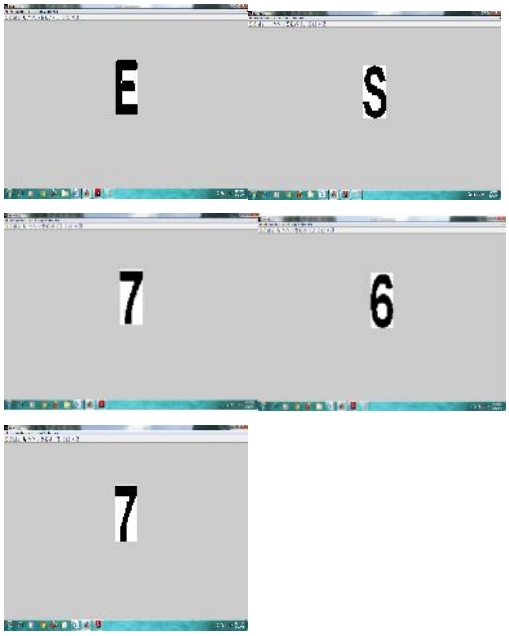



**Fig.5.10 Character segmented image of printed text**

### 5.5 Results of Recognition

In the recognition process, each character extracted is correlated with each and every other character present in the database. The recognized characters obtained are shown in the mat lab and notepad which are the final outputs as shown in below figures:

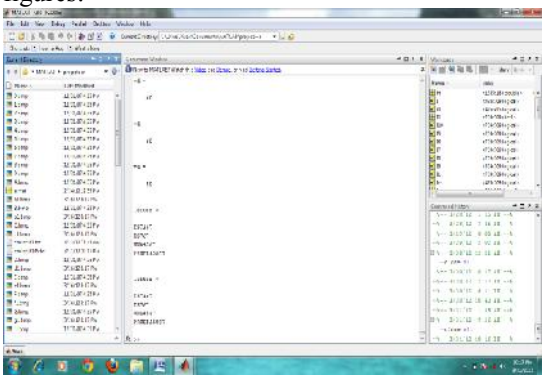


**Fig.5.11 Recognized output in Mat lab**

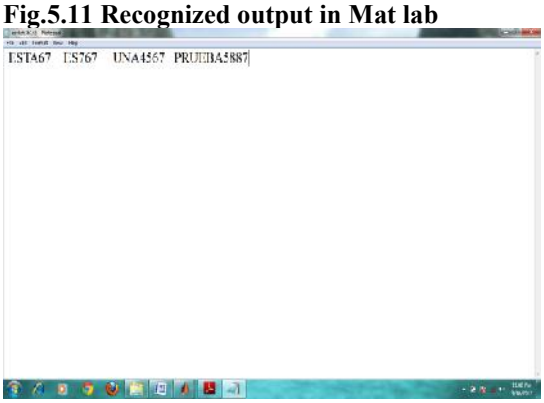



**Fig.5.12 Recognized output in notepad**

## I. CONCLUSION

Recognition of characters in document images using morphological operationtechnique is easy to implement. Since this algorithm is based on simple correlation with the database, the time of evaluation is very less. Also the database which was partitioned based on the areas of the characters made it more efficient. Thus, this algorithm provides an overall performance in both speed and accuracy. “Optical Character Recognition” using correlation, works effectively for certain fonts of English printed characters. This has applications such as in license plate recognition system, text to speech converters, postal departments etc. It also works for discrete handwritten run-on characters which has wide applications such as in postal services, in offices such as bank, sales tax, railway, embassy, etc.

Since “Character Recognition” deals with offline process, it requires some time to compute the results and hence it is not real time. Also if the handwritten characters are connected then some errors will be introduced during the recognition

process. Hence the future work includes this to be implemented for an online system. Also this has to be modified so that it works for both discrete and continuous handwritten characters simultaneously.

**REFERENCES**

[1] H.K. Kwag, S.H. Kim, S.H. Jeong, G.S. Lee, Efficient skew estimation and correction algorithm for document images, Image and VisionComputing 20 (2002) 25–35.
[2] B. Yu, A.K. Jain, A robust and fast skew detection algorithm for generic documents, Pattern Recognition 29 (10) (1996) 1599–1629.
[3] E. Kavallieratou, N. Fakotakis, G. Kokkinakis, Skew angle estimation for printed and handwritten documents using the Wigner–Ville distribution, Image and Vision Computing 20 (2002) 813–824.
[4] L. O'Gorman, The document spectrum for page layout analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 15 (11) (1993) 1162–1173.
[5] R.M. Haralick, Monocular vision using inverse perspective projection geometry: analytic relations, Proceedings of the IEEE ComputerVision and Pattern Recognition Conference 1989; 370–378.
[6] P. Clark, M. Mirmehdi, Recognizing text in real scenes, International Journal of Document Analysis and Recognition 4 (4) (2002) 243–257.
[7] M. Pilu, Extraction of illusory linear clues in perspectively skewed documents, Proceedings of the IEEE Computer Vision and PatternRecognition Conference 2001; 363–368.
[8] C.R. Dance, Perspective estimation for document images, Proceedings of the SPIE Conference on Document Recognition and RetrievalIX 2002; 244–254.
[9] P. Clark, M. Mirmhedi, Rectifying perspective views of text in 3Dscenes using vanishing points, Pattern Recognition 36 (2003)2673–2686.
[10] W.B. Irving, Applied Statistical Methods, Academic Press, New York, 1974.
[11] R. Jain, R. Kasturi, B.G. Schunck, Machine Vision, McGraw-Hill, New York, 1995.
[12] R. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, Cambridge, 2000.
[13] Y. Yang, H. Yan, An adaptive logical method for binarization of degraded document images, Pattern Recognition 33 (2000) 787–807.
[14] Y. Liu, S.N. Srihari, Document image binarization based on texture features, IEEE Transactions on Pattern Recognition and MachineIntelligence 19 (1997) 540–544.
[15] L. O'Gorman, Binarization and multithresholding of document images using connectivity, Graphical Models and Image Processing56 (1994) 496–506.
[16] N. Otsu, A threshold selection method from grey-level histograms, IEEE Transactions on System, Man, Cybernetics SMC-8 (1978)62–66.
[17] P. Soille, Morphological Image Analysis: Principles and Applications, second ed., Springer Verlag, Berlin, 2003.
[18] L.A. Zadeh, Calculus of Fuzzy Restrictions, Fuzzy Sets and Their Application to Cognitive and Decision Making Processes, AcademicPress, New York, 1975.

Mr. K. Ravi Kumar ,Received M.Tech (CSE) from PragatiEnggCollege,JNTU Kakinada,working as Head of the Department of CSE, Kakinada Institute of Engineering and Technology, Korangi, Kakinada. He has 7 years of teaching experience. He has published 07 papers in both National & International Journals. His area of Interest includes Bio-Informatics, image processing and Data Mining. He guided many projects for Postgraduates and Graduates.

Dr. G. Lavanya Devi, Asst.Prof, College of Engineering Andhra University, Visakhapatnam. She is working as Asst.Prof in Dept of CSSE in Andhra University. She has 10 Years of experience in teaching . She received her Ph.D from JNTUHyderbad on Bio- informatics. She published 09 National papers& 05 International Journals. She was guiding 05 research scholars in Bioinformatics and image processing.

Mr.P. EswaraBabu is a student of, Kakinada Institue of Engineering and Technology, Korangi, Kakinada. Presently he is pursuing his M.Tech (C.S) from JNTUK and he received his Bachelor of Technology (B.Tech) in Information Technology from JNTU, Kakinada in the year 2006. His area of interest includes Computer Networks, Data Mining and Database Management Systems.