4 PREPROCESSING

One of the objectives of this research is to segment Gujarati handwritten text document. The output of OGHTR's segmentation phase is individual segmented unit which is then further process for feature extraction. But the acquired scanned image data tends to produce different types of noises which causes improper segmentation and feature extraction.

The text recognition application uses preprocessing techniques to simplify the task of recognition system, organizing the information and representing it in a more accessible manner so that it gives reliable results. It deals with not only noise removal from the document image but also for reduction in data and improving quality for better processing [30,46].

Hence the preprocessing stage concerned with processing of scanned images to make it suitable for segmentation and feature extraction so that segmented unit exhibits features of the character which contributes in recognition process [6,52,106,120,126]. The preprocessing steps consist of sub steps like conversion of image from colour to gray or gray to binary; reduce noise present in the image data, contrast enhancement, stroke normalization, skew detection and correction.

As discussed earlier in section 1.4 the subset of preprocessing step selection is based on application, presence of noise and to reduce processing. Here in our research we have created our own dataset for experiment. The text is collected on A4 white paper and scanned on flat-bed scanner.

For our pattern the preprocessing activities includes grayscale conversion, contrast normalization, noise removal, binarization, skew detection and correction, size normalization and thinning etc. The preprocessing steps followed in IGHTR system shown in Figure 12 to make input document image suitable for segmentation.

The input to the preprocessing stage is physical document and output is binary document which is free from noise, document skew and cropped to optimum area.

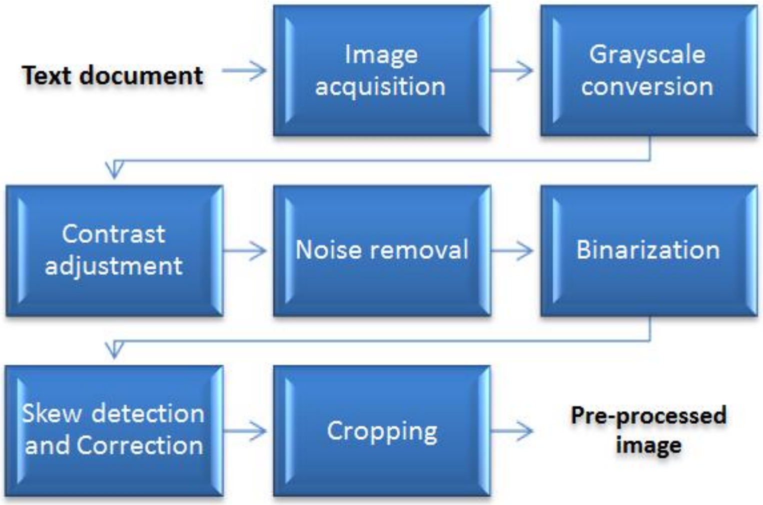


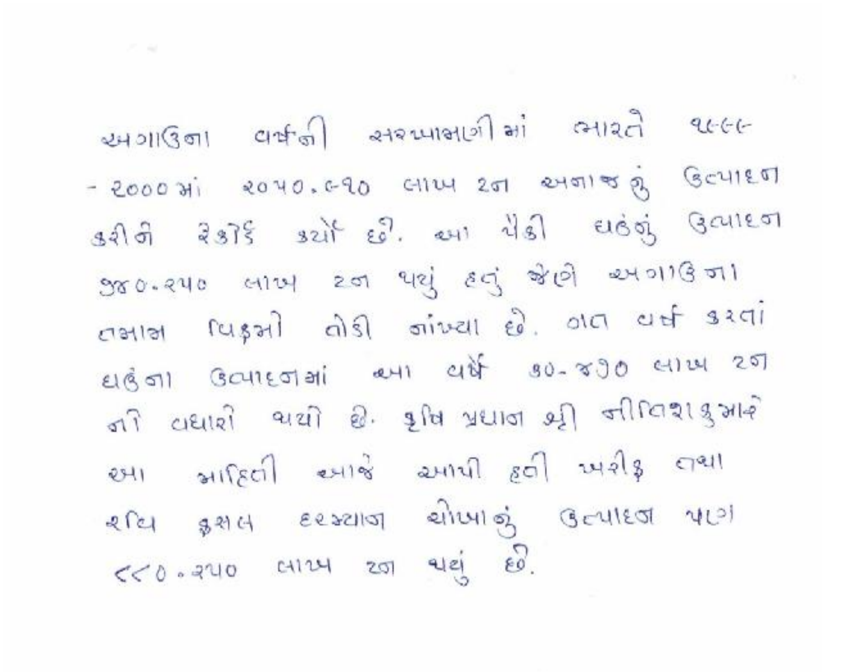
Figure 12 Preprocessing steps in OGHTR

4.1 IMAGE ACQUISITION

The main objective of this step is to covert hard copy document into electronic format. The offline handwritten document is converted to digital form using scanner or camera to produce offline text image. The handwritten document is fed into the flatbed scanner for digitization is widely used technique. While camera is used in condition where paper is fragile and cannot be forced to be flat.

The input image captured through camera need different process than scanned document image. The scanner is widely used device to digitize text paper document. While scanning process selection of appropriate resolution plays an important role in document quality.

The text document images were scan with 300 dpi resolution using HP flat-bed scanner. Below Figure 13 shows a scanned document image used in our experiment. The different types of noises appears in scanned document due to scanning process [33,34,61,69,92,100,113]. The image scanned in colour mode and stored in jpg format.



અગાઉના વર્ષની સરખામણીમાં ભારતે ૧૯૯૯
- ૨૦૦૦માં ૨૦૫૦.૯૧૦ લાખ ટન અનાજનું ઉત્પાદન
કરીને રેકોર્ડ કર્યો છે. આ પૈકી ઘઉંનું ઉત્પાદન
૭૪૦.૨૫૦ લાખ ટન થયું હતું જેણે અગાઉના
તમામ વિક્રમો તોડી નાંખ્યા છે. ગત વર્ષ કરતાં
ઘઉંના ઉત્પાદનમાં આ વર્ષે ૩૦-૪૭૦ લાખ ટન
નો વધારો થયો છે. કૃષિ પ્રધાન શ્રી નીતિશકુમારે
આ માહિતી આજે આપી હતી ખરીફ તથા
રવિ ફસલ દરમ્યાન ચોખાનું ઉત્પાદન પણ
૮૮૦.૨૫૦ લાખ ટન થયું છે.

Figure 13 A scanned document image

4.2 GRAYSCALE CONVERSION

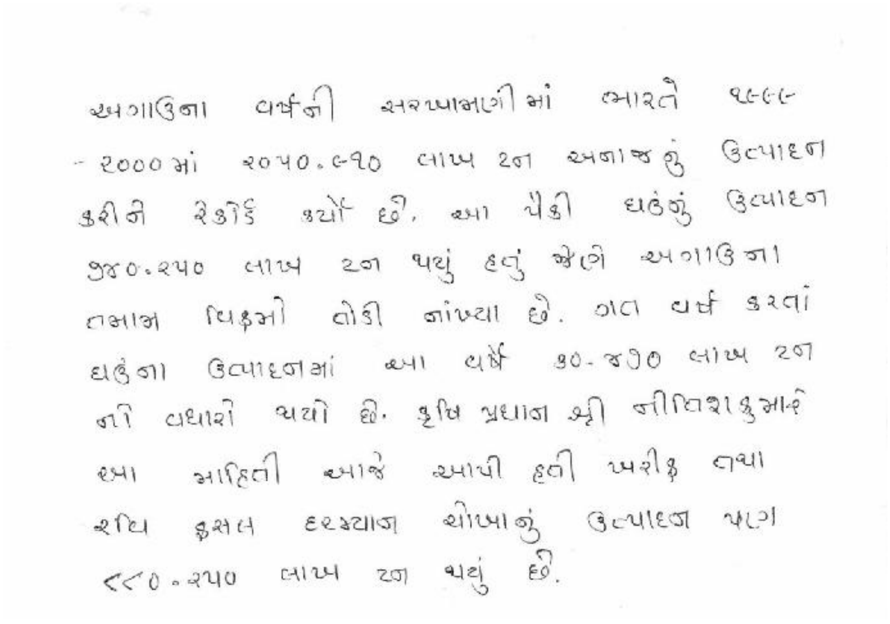
Preprocessing steps Grayscale conversion of OGHTR deals with converting colour image (RGB image) document into grayscale. The text in document image is best represented as binary image, black background with white trace of characters. The conversion to binary is greatly enhanced when the scanned document originates as colour. Scanning document in colour (RGB) captures much greater degree of detail than simple black and white.

Even grayscale has higher tonal values captured during scanning than black and white. Due to that we scanned document image in colour. To convert an image into two-dimension plane, i.e. binary image, due to noises it cannot be converted directly from colour image to binary image. Also, we want to improve quality of edges which represent the text. Due to that before converting image into binary first colour image is converted into grayscale image.

The conversion of image from RGB to grayscale is done by eliminating the hue and saturation information while retaining the luminance. To convert colour image to grayscale image using Equation 2. The equation uses weighted sum of the R, G, and B components. The weightage is based on how the human eye perceives the red, green and blue colours [65].

$$g(x,y) = 0.2989 * f(x,y,1) + 0.5870 * f(x,y,2) + 0.1140 * f(x,y,3) \qquad \text{Equation 2}$$

Here in this equation g refers to the output grayscale image and f represents input true colour image. In input image f, value of 3rd parameter 1, 2 and 3 refers to red, green and blue colours.



અગાઉના વર્ષની સરખામણીમાં ભારતે ૧૯૯૯
- ૨૦૦૦માં ૨૦૫૦.૯૧૦ લાખ ટન અનાજનું ઉત્પાદન
કરીને રેકોર્ડ કર્યો છે. આ પૈકી ઘઉંનું ઉત્પાદન
૭૪૦.૨૫૦ લાખ ટન થયું હતું જેણે અગાઉના
તમામ વિક્રમો તોડી નાંખ્યા છે. ગત વર્ષ કરતાં
ઘઉંના ઉત્પાદનમાં આ વર્ષે ૩૦.૪૭૦ લાખ ટન
નો વધારો થયો છે. કૃષિ પ્રધાન શ્રી નીતિશકુમારે
આ માહિતી આજે આપી હતી ખરીફ તથા
રવિ ફસલ દરમ્યાન ચોખાનું ઉત્પાદન પણ
૮૮૦.૨૫૦ લાખ ટન થયું છે.

Figure 14 Text document converted to grayscale

The document image after converting into grayscale is shown in Figure 14. After converting an image into grayscale, image quality is enhanced using next preprocessing steps. These steps are contrast enhancement and noise removal.

4.3 CONTRAST ADJUSTMENT

The handwritten text within the documents often shows certain amount of variation in terms of the stroke width, stroke brightness, stroke connection and document background. These tend to induce the thresholding problem for binarization due to high inter-intra-variation between the text stroke and document background [120].

The contrast adjustment enhances document image by making dark portion darker and bright portion brighter. It can be enhanced using contrast stretching or histogram equalization method [65]. A contrast stretching is a simple image enhancement technique that attempts to improve the contrast in an image by stretching the range of intensity values so it contains to span a desired range of values [65]. It is a linear scaling function to the image pixel values. As a result the enhancement is less harsh.

The more sophisticated method is histogram equalization, spreads out intensity values along the total range of values to achieve higher contrast [65]. A modified contrast limited adaptive histogram equalization (CLAHE) method used to adjust intensity of document image [127]. The method divides entire image into sub-image and histogram is applied to all small sub-images like in adaptive histogram equalization and then result is interpolated.

4.4 NOISE REMOVAL

The output of contrast adjustment is grayscale image with enhanced contrast in its pixels value. This grayscale document is then processed for noise removal. The objective noise removal step is to remove unwanted bit pattern which do not contribute in text recognition process and it is considered as noise.

The noise refers to the error in the pixel value or an unwanted bit pattern, which do not have any significance in an output. The scanner, a most popular offline acquisition device, introduces imperfections of its own. The noise introduced by the optical scanning device or the writing instrument like pen, causes disconnected line segments, bumps and gaps in lines, filled loops, isolated point etc.

There are many techniques available in literature to reduce the effect of noise based on filtering and morphology. The morphological operation is applied on binary image. The filtering based method which are used for noise removal are smoothing, sharpening, median and Wiener filter [6,65,102,128].

We used median filter for noise removal as the median filter is the effective way of reducing salt and pepper noise caused by the document scanning process [65]. The main idea of the median filter is considers each pixel in the image in turn and looks at its nearby neighbours to decide whether or not it is representative of its surroundings. It replaces centre pixel value with the median of those values as shown in Figure 15.

The median is calculated by first sorting all the pixel values from the surrounding neighbourhood into numerical order and then replacing the pixel being considered with the middle pixel value [65].

140	139	138	127	139
139	138	134	123	125
134	134	**80**	128	121
140	130	138	122	139
128	120	139	131	136

Pixels values from 3 X 3 neighborhood:

138, 134, 123, 134, 127, 128, 130, 138, 122

Median value is: **130**

Figure 15 The process of Median filter

At this stage noise in the image is removed which introduced during scanning process. The unwanted bit-patterns which are less than the size of mask are removed from the image. The noises which are bigger than the mask size and outside of text area will be taken care in cropping step in 4.6. Now image is ready for binarization step.

4.5 BINARIZATION

Once the image is free from noise it is processed for binarization. As we discuss earlier that our interest lies in text region of document image. It is better to represent text as foreground and page as background. The input to binarization step in preprocessing is grayscale image and output is binary image.

The binarization step will convert whole document into binary and reduces data which required to be processed. To convert gray image into binary image thresholding technique is used.

Binarization step is accomplished by scanning the image pixel by pixel and assigning each pixel as text or background, depending on whether the grey level is greater or lesser than the value of threshold T [65]. The Equation 3 is used for converting image into binary is:

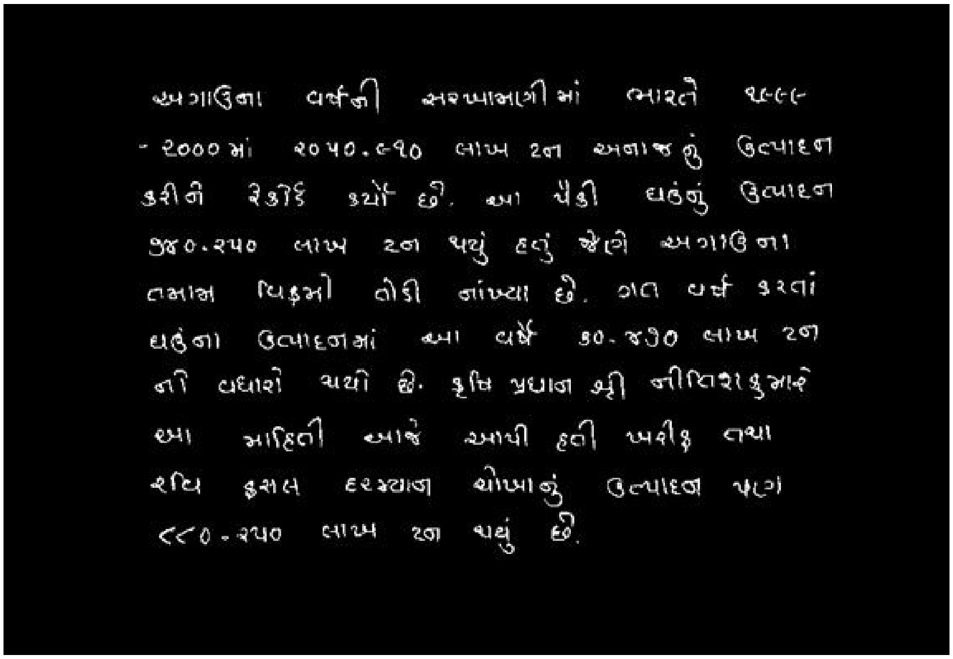
than the val

into binary s:)

$$g(x,y) = \begin{cases} 1 & f(x,y) < T \\ & f(x,y) \geq T \end{cases}$$

Equation 3

Choosing single threshold value for whole document is also known as global threshold. Picking up the threshold value can be done by human operator, by calculating mean gray level, and by analysing the histogram of an image. Otsu's thresholding is used by many researchers [10,54,56].



અગાઉના વર્ષની સરખામણીમાં ભારતે ૧૯૯૯
- ૨૦૦૦ માં ૨૦૫૦.૯૧૦ લાખ ટન અનાજનું ઉત્પાદન
કરીને રેકોર્ડ કર્યો છે. આ પૈકી ઘઉંનું ઉત્પાદન
૭૪૦.૨૫૦ લાખ ટન થયું હતું જેણે અગાઉના
તમામ વિક્રમો તોડી નાંખ્યા છે. ગત વર્ષ કરતાં
ઘઉંના ઉત્પાદનમાં આ વર્ષે ૬૦.૪૭૦ લાખ ટન
નો વધારો થયો છે. કૃષિ પ્રધાન શ્રી નીતિશકુમારે
આ માહિતી આજે આપી હતી ખરીફ તથા
રવિ ફસલ દરમ્યાન ચોખાનું ઉત્પાદન પણ
૮૮૦.૨૫૦ લાખ ટન થયું છે.

Figure 16 Binarized handwritten document

Otsu binarization is a global thresholding method with threshold fixed based on minimization of the weighted sum of within-class variances of the foreground and background pixels using gray level histograms [65,68].

For better processing and convenience the image is inversed so that 1 represent as foreground and 0 as background. The result of binarization step is shown in Figure 16. Note that we must have to complement the binary image to represent foreground in 1s and background in 0s.

The morphological operations on binary image are very much useful to improve image data [65,100]. Morphological cleaning is applied to remove isolated pixels from the binary image that will removes object pixel 1s which is surrounded by 0's. Morphological dilation is performed to fill the gap and increases the foreground [65].

4.6 CROPPING

After the binarization image is ready for segmentation but document image may have empty space at top, bottom, left and right of text region. This extra information in the binary image does not contribute in further process. Hence it is to be removing using cropping of OGHTR preprocessing step.

Easiest way to crop document image up to available text area is using extracting bounding box that contain text area. The process of extracting bounding box is to get maxima and minima in X direction, and maxima and minima in Y direction. Using these parameters rectangle area containing text region is extracted from the image.

Even after applying the above cropping process to binary document image, many of document images did not cropped to text region. It contains empty spaces which do not have text lines. This is due to noise caused by small group of isolated pixels in the document. The binary handwritten text document has this isolated small group of pixels, due to touching of pen to the paper while writing, paper quality or scanning errors. It is not representing any of the part of handwritten text so considered as noise.

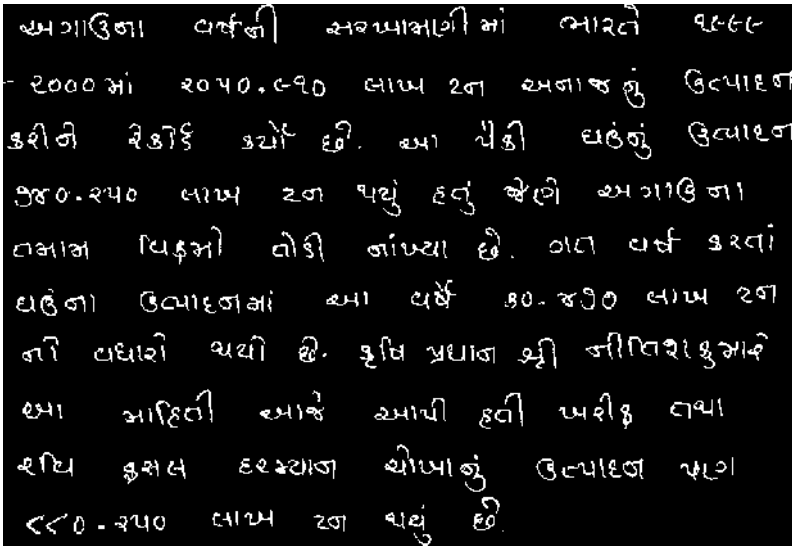
The reason behind the isolated group of pixels are not removed during noise removal step is due to mask size used for applying median filter. It cannot eliminate noises which are greater than the size of mask [65]. Choosing mask more than the stroke width pixel (edge width) causes broken edges in the character image.

Morphological erosion can be used to remove this isolated group of pixels with structuring element size bigger than group size [65]. But if we perform morphological erosion on whole text image it also erodes the textual part result in broken edges as well as it removes some of the important object like *anusvār* (am “◌□”) in Gujarati text.

To solve this problem, the copy of binary image is taken and morphological erosion is performed with square structuring element. The size of structuring element is based on stroke width. The discussion on finding stroke width is given in section 5.1. Using eroded binary image horizontal and vertical projection is computed and parameters for optimum text area obtained using extracting bounding box which contains text area. The original binary images then shrink to available text area using these parameters. The steps to perform cropping is as given below:

Algorithm:	**Cropping**
Input:	Binary Image
Output:	Cropped binary image
1.	Apply erosion to binary image using square structuring element of size 3 X 3.
2.	Find minimum and maximum row and column of object pixel as minrow, maxrow, mincol and maxcol
3.	Extract region covered by rectangle (minrow,mincol) to (maxrow,maxcol)
4.	END

After applying cropping binary document image is transform to the image as shown in Figure 17. The cropped binary image may suffer from skew so it is processed for skew detection and correction in next step.



અગાઉના વર્ષની સરખામણીમાં ભારતે ૧૯૯૯
-૨૦૦૦માં ૨૦૫૦.૯૧૦ લાખ ટન અનાજનું ઉત્પાદન
કરીને રેકોર્ડ કર્યો છે. આ પૈકી ઘઉંનું ઉત્પાદન
૭૪૦.૨૫૦ લાખ ટન થયું હતું જેણે અગાઉના
તમામ વિક્રમો તોડી નાંખ્યા છે. ગત વર્ષ કરતાં
ઘઉંના ઉત્પાદનમાં આ વર્ષે ૬૦.૪૭૦ લાખ ટન
નો વધારો થયો છે. કૃષિ પ્રધાન શ્રી નીતિશકુમારે
આ માહિતી આજે આપી હતી ખરીફ તથા
રવિ ફસલ દરમ્યાન ચોખાનું ઉત્પાદન પણ
૮૯૦.૨૫૦ લાખ ટન થયું છે.

Figure 17 Binary cropped document

4.7 SKEW DETECTION AND CORRECTION

Skew detection and correction are used to align the paper document with the coordinate system of the scanner. Due to the improper alignment setting during the scanning process the paper may be tilted in either of the direction which causes the rotation of image. Sometimes in handwritten document improper writing fashion leads to rotation. Main approaches for skew detection include correlation, projection profiles, Hough transform, and linear regression [65,69,129].

o ind line reg ession [65,69,129]

$$x' = x * \cos\theta - y * \sin\theta$$
$$y' = y * \cos\theta + x * \sin\theta$$

Equation 4

Simple yet robust method for detecting skew angle is through histogram projection. The image is rotated using Equation 4, in clockwise direction with an angle θ and horizontal projection length is calculated. If projection length increases previous angle is noted. The process is repeated for anti-clockwise direction and angle is noted. The higher value of angle from clockwise and anti-clockwise direction represents skew in document.

Choosing small θ angle increasing computation as it is processing for rotating image. To save processing initial increment to the angle is taken as 5° and after getting skew angle increment value is taken as 1˚ and processed for 1˚ to 4˚ degree. The algorithm steps for skew correction and detection is given as below. It is divided into two routines. The SkewDetection routine is used to get rotation angle in clock wise and anti-clock wise direction. The SkewCorrection routine is used to compare the angle and rotate the image for skew correction.

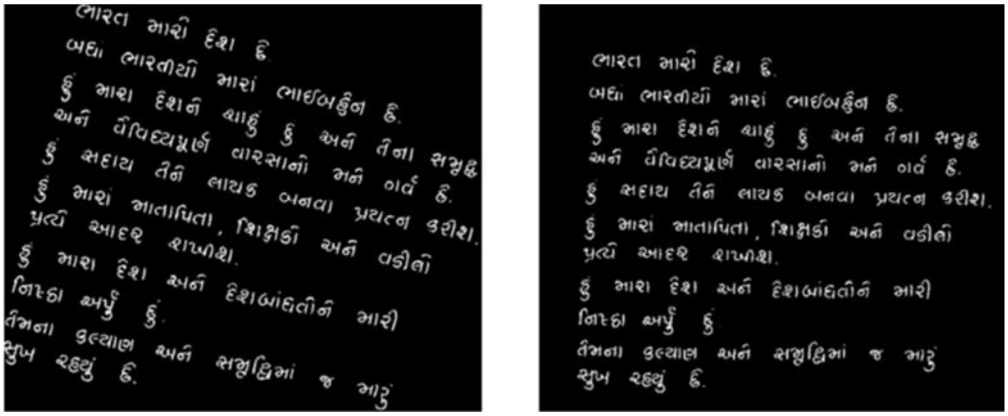
Algorithm:	**SkewCorrection**
Input:	Binary Image
Output:	Binary Image with skew correction

```
1.    ang1=SkewDetection(img, 1)
2.    ang2=SkewDetection(img, -1)
3.    If(ang1 > ang2)
          Rotate image with ang1
      ELSE
          Rotate image with -ang2
      END IF
4.    END
```

Algorithm:	**SkewDetection**
Input:	Binary Image, flag
Output:	Skew angle

```
1.    Calculate projection length of image
2.    FOR ang=5 to 90 STEP 5
3.        Calculate projection length after rotating image by ang*flag
4.        IF projection length increases on rotation
              ang = ang – 5*flag
                  Repeat process for 1 to 4 angle
              break;
          ELSE
              Replace projection length with rotated length
          ENDIF
5.    NEXT
6.    RETURN ang
```

After finding skew angle, image is rotated using Equation 4 to correct the skew. To experiment result of skew correction a document with skew is fed into the system and obtained result is shown in Figure 18.



ભારત મારો દેશ છે.
બધા ભારતીયો મારાં ભાઈબહેન છે.
હું મારા દેશને ચાહું છું અને તેના સમૃદ્ધ
અને વૈવિધ્યપૂર્ણ વારસાનો મને ગર્વ છે.
હું સદાય તેને લાયક બનવા પ્રયત્ન કરીશ.
હું મારાં માતાપિતા, શિક્ષકો અને વડીલો
પ્રત્યે આદર રાખીશ.
હું મારા દેશ અને દેશબાંધવોને મારી
નિષ્ઠા અર્પું છું.
તેમના કલ્યાણ અને સમૃદ્ધિમાં જ મારું
સુખ રહ્યું છે.

Figure 18 Document image before and after skew correction

Due to the rotation of image it may contain empty space as shown in Figure 18. So it need to again process for cropping. Now this time cropping is done using simple method by extracting bounding box region after finding minima and maxima.

4.8 DISCUSSION

In this chapter we explained the preprocessing of input text document image so that it is prepared for next segmentation phase. After the text document is digitized into RGB, it is passed to next steps grayscale conversion to convert RGB image to grayscale. The contrast adjustment step is to enhance image contrast. The noise removal step is performed to remove unwanted bit-pattern from an image which is noise using median filter. The binarization step is to convert grayscale image into binary i.e. two colour image. The binarization step reduced data of pixel to 1 bit representation.

Cropping is the process to shrink image to available text area. Our cropping method is based on morphological erosion operation and projection histogram which result in crop the binary image which take care of noise caused by isolated group of pixel and without losing information in textual data area. The final step of preprocessing is to correct document skew. The improper page alignment caused skew during scanning process is removed using skew detection and correction method using projection profile. We have applied said preprocessing stage on our dataset of handwritten text documents. We observed that the noise which is larger than the size of stroke width was not removed as it is available within text area. But if we process it for removal may cause damage to edges of the character so we decided to not to process.

Due to our main focus is toward segmentation and feature extraction we kept only essential steps in preprocessing of document. After applying preprocessing on the scanned image, it is ready to process for which will be discussed in next chapter.