5.1 Preprocessing

Preprocessing is the first step of any pattern recognition system. In case of character recognition, preprocessing is useful to reduce various type of noise from scanned document image. It also normalizes image and highlights the components of input image which are required for segmentation and feature extraction operations. It is considered that after applying preprocessing algorithm on input image, it must convert an image into more suitable format than the original image. For different purpose, a different type, of preprocessing are needed. Basically preprocessing is very much problem oriented and based on application type the preprocessing methods can be chosen. For example a preprocessing algorithm which is perfect for signature recognition may not be suitable for the purpose of face detection and recognition. Every image processing application requires specific type of feature which may be different from other system’s requirements. The primary objective of preprocessing in any digital image processing application is to make an image appropriate for feature
extraction. Similarly character recognition system needs specific feature array to be extracted from document image.

During document image acquisition, different types of noise become as a part of image and create difficulties during feature extraction due to optical scanning devices and poor document pages. To improve the accuracy of the character recognition system such type of noises must be removed in preprocessing step. Also document image may be tilted and skewness of certain angle come out in it and occurrence of skewness in document image may lead to unsatisfactory results of segmentation and feature extraction. By applying appropriate skew detection and correction methods such skewness and slants needs to be removed from the document image. During preprocessing, Normalization is required because handwritten document consists of a lot of handwriting variation. Similarly the printed document image may consist of different font type and size. Such variations need to be removed and data is standardized by applying various normalization techniques. Reducing data size is another objective of preprocessing, because any document image consists of very large number of pixels. For feature extraction all these pixels are not required. The pixels which provide information of character features are important. Due to this reason the compression techniques can be applied, which preserve shape information and remove other pixels from the image. It helps to reduce the storage requirement and enhance the processing of the system. It also improves the efficiency and accuracy of the system.

As preprocessing is specific for character recognition system, it requires specific document preprocessing where important information should not be removed. For any character recognition preprocessing should be robust, because removing a single dot may change the meaning of character. For example if a dot of character ‘sey’ (۸) is mistakenly removed with noise then the character will becomes ‘tey’ (۸). Similarly while preprocessing if two dot of ‘Cheem’ (۸) is removed then it will become Jeem (۸), if again dot is removed then it will become ‘Hay’ (۸). These characters can be differentiated only by the number of dots and their position. It means that secondary strokes like dots are very important part of the characters, so it is very necessary to perform preprocessing very carefully. In Character recognition of Urdu text, its basic aim is to remove the noise, normalize image against skewness,
slant, size etc. It is also considered that the storage requirement should be reduced, which will increase the processing speed.

5.1.1 Earlier Preprocessing Techniques in Urdu.

Due to its importance and the further results of feature extraction and recognition are strongly depended on preprocessing many researchers tried to introduce better preprocessing techniques and proposed various methodologies. [1] R.A Peter presented a new method by using mathematical morphology for noise reduction. In his paper he described new morphological image cleaning algorithm (MIC). This concept is helpful in case of gray-scale images which are corrupted by random noise, low-amplitude or patter noise. Also MIC is useful to conserve the thinning features during noise removal and image enhancement process. In this method the residual image is recombined with processed residual images. He discussed the results of noise removal with different variation effects. Imran Razzak [2] proposed a preprocessing method in which the input strokes of both offline and online are considered for obtaining detail information for preprocessing. They basically concentrate on smoothing, de-hooking, interpolation etc. Better results are reported by using parallel preprocessing by combining offline and online method. M. Hussain [3] used smoothing technique and computes the displacement between two points using displacement from 4,8,16.

Sheikh Faisal Rashid et. al. [4] Presented an approach for detecting orientation of Urdu documents images by different fonts and layouts using discriminative learning method. The method is based on classification of single connected components orientation present in a document image. Afterward the majority of count is used to determine page orientation. Also they used heuristic rules for removing marginal noise. The variation in shape orientation may occur due to changes in orientation but proposed system also able to learn shape variation. This ability to learn the variation in shape is useful in case of Urdu character because the majority of the Urdu letters consists of diacritics and dots which are nothing but small connected components with similar shape orientation. Due to this small connected components are also removed from document image which is drawback of this system. Malik et. al [5] presents an approach for Handwritten Urdu Word Spotting for this they used the method called connected components analysis.
They apply average filter on grayscale handwritten Urdu document image and it is converted into binary image. After conversion of image from grayscale to binary perform labeling on connected components. In [6] Saqib et. al. presented Urdu document image layout analysis, it is a robust system under various document image-degradation. They used multi-resolution morphology based method for separating text and non-text. Ridge based text-line finding method is used which is based on ridge detection and Gaussian filter bank smoothing. Their system can work for different languages like Urdu, Arabic, and Persian etc. [7] Used both grayscale and binarized images during preprocessing stage. After binarization for smoothing and reducing the salt and pepper type noise they applied Median filters on binarized document images. After smoothing and noise removal the bounding box method preferred to eliminate unwanted white space. They used aspect ratio technique and normalized image into two sizes 64 X 64 and 128 X 128. [8] S. Sardar et. al. proposed character recognition system for Urdu script in which they used gray scale thresholding for removing noise and converted image into binary. Somaya et. al. [9] performed normalization of strokes, slants, slope height and width of the letters. In their work they used algorithms proposed by [10] for baseline correction.

Safwan [11] used median filter and minimum filter of special for preprocessing and removing jaggedness of the word and character contours. The condition for applying these filters is that not remove a pixel considered to connect more than one blobs and prevents from loss of original data. Many researchers proposed Arabic and Urdu character recognition system which are robust to noise [12]. Present character recognition system which is robust to noise. One more character recognition system in [13] is also robust to noise and scale invariant. In this a new algorithm for optimized boundary representation is used. In languages like Arabic, Urdu, Persian and Jawi the dots are very important. The change in place or removal of dot may alter the meaning. In [14] author proposed a preprocessing technique for Persian document which removes the noise similar to the dot and diacritic. In this work after segmentation the dot size is estimated for each region. Based on the estimated side of these dots the noise is removed which is similar to the dots present in document image. The smallest rectangle for each connected component is specified. After specifying connect component smallest rectangle threshold is used to remove the noise components according to the estimated dot size.
In [15] the threshold is determined by modified version of maximum entropy sum and entropic correlation method and it is used for binarization. By using 3X3 structuring element; Morphological closing flowed by morphological opening is performed to remove spurious segments. For image compression binarized document image is surrounded using circumferential rectangular around exact image. [16] Used explicit noise model in which these noise models are trainable automatically. In this, it is assumed that mostly the noise is present close to the boundary of word and intensity of noise reduces with rising distance from boundary. The author in [17] Suppress small island type noise which don’t contribute shape information, block connectors and narrow channels is reduced using closing operation followed by opening operation come under morphological preprocessing. In [18] for noise reduction image is converted into monochrome image from which the noise element are removed.

In [19] A. Cheung used Otus method for document binarization. After binarization smoothing technique is applied to minimize the noise appeared due to shading effect. For Arabic and Persian document noise removal, Mandana K et. al. [20] used word and sub-word method with two threshold values TN and TM. In this method the words and sub-words are extracted and their contours are classified by analyzing it. The contour’s class is found by setting threshold value TN and TM. Contour’s length is defined by CL and considered that if the value of TN > CL this contour is considered as noise and are set as 0. On the other hand when CL is greater than or equal to TM the contour may considered as the main body of the word or sub-word and the field values are set as 1. In this way the noise are set zero and main body are of word becomes noise free. In [21] a noise removal system based on median filter is used for Arabic character recognition. In this method they used Median filters of 3X3 and 5X5 size windows. The noise removal using these filters was observed satisfactory for Arabic characters. [22] Mentioned linguistic topology rules and shows that diacritics and dots are noise sensitive. These dots may be placed below or above of the main body of script, in some scripts these are not placed below or above the characters. Also enlighten character candidates as combination of classified scripts which are associated through classified dots. In [23] smoothing followed by thresholding is used for reducing noise from document image. After noise reduction the thinning is performed reduce pixel representing characters. Mainly the salt and pepper noise is considered while removing the noise from the document image.
In [24] region growing technique is used for noise removal from document image. The linear regression method is also used for removing the role lines printed on pages. Specifically to remove the pepper noise the text pivoting technique is implemented. Abuhaiba et. al. in [25] applied CBSA algorithm and noted that the CBSA (Clustering based Skeletonization Algorithm) is better to tolerate the document image noise as compared to other algorithms for thinning. Usually a CBSA method considers a character image as a set of adjacent clusters which provide sufficient information to represent character. In [26] Abdullah et. al. for rejecting image noise the coefficients are used which comprise as large magnitude as possible. For normalization of scale each coefficient is divided by their specific magnitude. The Jawi script is also similar to Urdu script so the noise removal techniques used for Jawi may also word better for noise removal of Urdu characters. Mohammad Faidzul Nasrudin et. al [27] presented a fruitful review on Jawi handwritten character recognition and summarized many techniques for noise removal in Jawi script. In their work they review various proposed methods [31-39] for preprocessing and normalization. Khairuddin et.al. [28] used a sequence of preprocessing methods in which first Median filter is used to reduce the noise present in a Jawi document image then binary image is produce by thresholding technique as proposed in [29]. After image enhancement the Gradient orientation histogram was implemented to normalize the slants and skew present in document. Finally the image compression is done using thinning algorithm of [30].

For Jawi handwritten scrip Mazani Manaf M. [31] detected color of the digitized document image using gamma and intensity correction as suggested by[32] and converted it grayscale by implementing linear function proposed in [32, 33]. Then for noise removal morphological filtering is applied in which erosion operation is followed by dilation known as morphological opening. Following this the Median filter is applied for noise removal which is one of the spatial domain techniques. After that thresholding is done using threshold value which was arithmetic mean of the gray values of background pixels. Then for slant and skew correction Least Square Root and Gradient Orientation Histogram are used respectively. And finally thinning algorithms proposed in [34] are applied for compression of an image. In [35] a 3X3 structuring element is used and morphological closing is performed to block the narrow channels plus slim lakes. Similarly smoothing is performed by applying
opening operation. Mohd Sansui [36] presented a reengineering method in which Gradient Orientation Histogram is used for slant and skew normalization, also 3X3 Sobel filter is applied on grayscale image. Nafiz et. al. [37] categorized and presents a summary of all the character recognition methods systematically. It gives a brief of all the steps involved in any character recognition system.

5.1.2  **Preprocessing Techniques in CR**

Preprocessing is very much problem oriented and the preprocessing methods can be chosen according to the type of application. For example a preprocessing algorithm which is perfect for signature recognition may not be suitable for the purpose of face detection and recognition. In case of character recognition specific type of preprocessing is required. The Figure 5.1 illustrates the basic objectives of preprocessing step in any character recognition system. In which preprocessing objectives for character recognition can be categorized into three groups’ noise reduction or Image enhancement, normalization, and compression [38].

![Figure 5.1 Primary objectives of Preprocessing in Character Recognition](image)

5.1.2.1  Noise Reduction

The first objective of preprocessing is noise reduction also called as Image Enhancement. While scanning noise of various types introduced due to poor quality of scanning device, inappropriate illumination, and variation may be encountered in scanned document image. These differences in document image like broken characters lead to wrong feature extraction results. Overprinted text needs to be normalized before feature extraction; otherwise this may leads to poor accuracy. In
case of handwritten character recognition, noise removing and normalization play important role, because handwritten document page consists of numerous character variations and heavy noise. De-noising and normalization may be more challenging in case of handwritten document. The better feature extraction is necessary to remove such imperfections form document images. Hundreds of available noise reduction techniques can be categorized in three major groups called Filtering, Morphological operation based filtering and Noise Modeling [38, 39, 40].

5.1.2.1.1 Filtering:

Image enhancement or noise removing is subjective and problem oriented task. In character recognition, many researchers preferred filtering technique for image enhancement. In literature numerous filtering techniques can be found which can be broadly categorized into two groups known as spatial domain filter and Frequency domain filters.

--- Spatial domain filter

The term spatial domain refers to the aggregate of pixels composing an image [41]. This type of filtering is based on convolution of an image using specific type of mask. Spatial domain filtering methods can be grouped into three categories called Point Processing Techniques; Mask processing techniques and Histogram based techniques. If a mask has neighborhood size of \(1 \times 1\), then such filtering technique is known as point processing techniques. Figure 5.2 (a) illustrate a mask of \(1 \times 1\) size. Similarly if neighborhood size is \(3\times3\), \(5\times5\) etc. then such operation is called as Mask operation or Spatial filtering. In case of Mask processing filter the sub-image is known as mask or filter shown in Figure 5.2 (b).

![Figure 5.2](image)
— **Frequency Domain Methods**

Even though spatial domain filtering techniques is easy to understand, implement but frequency domain method has few distinguishing benefits for preprocessing. The convolution and correlation are computationally can be performed in better and efficient way using frequency domain preprocessing methods. This method is computationally beneficial and reduces requirement of storage and bandwidth for preprocessing and other character recognition steps.

The One dimensional Fourier transform for a single variable \( f(x), x = 0, 1, 2 \ldots N - 1 \) discrete function is given by

\[
F(u) = \frac{1}{N} \sum_{x=0}^{N-1} f(x) e^{-j2\pi ux/N} \quad u = 0, 1, \ldots N - 1 \quad (1)
\]

The original function can be obtained using Discrete Fourier Transform DFT.

\[
F(x) = \sum_{u=0}^{N-1} F(u) e^{j2\pi ux/N} \quad x = 0, 1, \ldots N - 1 \quad (2)
\]

The corresponding discrete Fourier transform pair for the two dimensional function is

\[
F(u, v) = \frac{1}{MN} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) e^{-j2\pi (ux/M + vy/N)} \quad (3)
\]

\[
F(x, y) = \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} F(u, v) e^{j2\pi (ux/M + vy/N)} \quad x, y = 0, 1, \ldots N - 1 \quad (4)
\]

When an image is to be processed in the frequency domain following steps need to be followed [41].

— First multiply the digitized image by \((-1)^{x+y}\) to center the transform.

— Transform the resultant image into frequency domain by using Discrete Fourier Transform DFT i.e. \( F(u, v) \).

— Perform the process on \( F(u, v) \) using the filter function \( H(u, v) \), in frequency domain the process will be multiplication because the convolution in the frequency domain reduces to multiplication in spatial domain and vice versa i.e. \( f(x, y) h(x, y) \leftrightarrow F(u, v) * H(u, v) \).

— The complete process is carried out on image in frequency domain.
— After performing preprocessing operations inverse DFT of the result i.e. perform operation IFTT (Inverse Fourier transform) on image which transforms the image again into spatial domain.
— Finally multiply the result by $(-1)^{x+y}$.

![Figure 5.3 Frequency domain filtering methodology](image)

In Figure 5.3 diagram $f(x,y)$ is an original image and $f'(x,y)$ is the processed image. $H(u,v)$ is filter function for process used in transform domain, $G(u,v)$ is the result of preprocessing in frequency domain. It can be represented by

$$G(u,v) = F(u,v)H(u,v) \quad (5)$$

In frequency domain filters can be categorized into low-pass filters and high-pass filters.

5.1.2.1.2 **Morphological operation based filtering:**

The morphological techniques such as morphological filtering, thinning, and pruning are useful for preprocessing and image enhancement. Mathematical morphology is also important tool for extracting image components that are useful in the representation and description of region shape, such as boundaries, skeletons, etc. These filters can be used for thickening, thinning the images which are improves broken characters, removes unwanted pixels like noise etc. The Morphological filters are based on two operations dilation and erosion. All the algorithms for
morphological image filtering are based on dilation and erosion. Dilation is usually applied on binary image for thickening or growing the objects. On the other hand erosion operations are used for thinning or to shrinks the objects present in binary images. For both dilation and erosion operations a sub image is used which is known as structuring element. The structuring element is nothing but a matrix consists of 0s and 1s arranged in specific manner as per requirement of filtering. The structuring element gets translated to several locations of an image. In case of dilation the translation process is like spatial convolution.

5.1.2.1.3 Histogram equalization:

Histogram equalization automatically determines a transformation function that seeks to produce an output image that has a uniform histogram. When automatic enhancement is desired then histogram equalization is a good approach. It is a technique for adjusting image intensities to enhance contrast. In histogram equalization, the input pixel intensity, r is transformed to new intensity value, s by transformation function T.

5.1.3 Tested methods for noise reduction:

In present work various filtering techniques are tested to find the best one for Urdu word document. Various types of noisy Urdu images are used to verify the filtering techniques. Initially Spatial Low-pass filters like average filter, median filter, and Max filter were applied and tested on Urdu document. Also histogram equalization and frequency domain Gaussian Low-pass filter are used to enhance the Urdu document image. In addition averaging Filter has also tested for smoothing the image, it is a special filtering technique where value of a particular pixel (x,y) is the average (arithmetic mean) of all the neighboring pixels. The smoothing is used to reduce the sharp transitions in the gray levels. The edges in image are blurred after applying average filter on it. Figure 5.4 shows a 3 X 3 mask where the constant multiplication for mask is equal to sum of the values of its coefficients.

\[ g(x, y) = \frac{1}{9} \sum_{i=-1}^{1} \sum_{j=-1}^{1} f(x + i, y + j) \]  \hspace{1cm} (6)
Weighted Average Mask:

To avoid effect of blurring, another kind of mask known as weighted average mask was found useful. In this mask the pixel at central location gets the maximum weighted. And the weight decreases while moving away from the center of the mask like.

Median Filter:

The response of this filter depends on ordering of the intensity values of pixels in the neighborhood of the pixel which is under consideration. In this type of filtering, at first set of intensity values of all the neighborhood of (x,y) are taken. Then arrange all these intensity values in particular order. Based on these ordering, select median of these values and this will be the value of center pixel. The median filter is specially used for salt and paper noise.

Max Filter:

In Maximum filter each pixel of an image is replaced with the maximum value of neighborhood pixel. In this type of filter a mask like 3X3 is passed over the image and the central pixel of the image is assigned as the maximum value among the entire neighborhood the pixels.

Smoothing in Frequency-Domain (Low-pass filters):

The properties like brightness and texture of an image are associated with Low-frequencies. In any image the low-frequency information is useful to process the properties like brightness and texture of the images etc. Low-pass filters can be used for this purpose. According to the requirement and types of the applications following
Low-pass filters can be used for smoothing of image. All three Low-pass filters and their transfer function are given in Table 5-1.

**Table 5-1 Lowpass filters**

<table>
<thead>
<tr>
<th>Low pass filters</th>
<th>Transfer function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ideal Low-pass filters</td>
<td>$H(u, v) = \begin{cases} 1 &amp; \text{if } (u^2 + v^2) \leq D_0^2 \ 0 &amp; \text{otherwise} \end{cases}$</td>
</tr>
<tr>
<td>Butterworth Low-pass filters</td>
<td>$H(u, v) = \frac{1}{1 + \left(\frac{(u^2 + v^2)^{1/n}}{D_0^n}\right)}$</td>
</tr>
<tr>
<td>Gaussian Low-pass filters</td>
<td>$H(u, v) = e^{-D_0^2(u^2 + v^2)/2D_0^2}$</td>
</tr>
</tbody>
</table>

**Frequency domain Gaussian Low-pass filter:**

When an image is to be processed in the frequency domain, the digitized image will first multiply the by $(-1)^{x+y}$ to center the transform. After multiplication transform the resultant image into frequency domain by using Discrete Fourier Transform DFT. Then perform the process on it using filter function in frequency domain. It can be defined by

$$H(u, v) = e^{-D_0^2(u^2 + v^2)/2D_0^2} \quad (7)$$

Where $D_0$ is cutoff of frequency

5.1.4 **Proposed Noise reduction Method:**

In proposed system, noisy documents are categorized into two groups. First types of documents are very dark and noisy document which are difficult to read and understand even by human eyes. These document images are produced due to scanning devices of small dynamic range, very dark page background and poor document pages etc. shown in Figure 5.5 (a). For such document image Algorithm-I is proposed. Second category of document consists of those documents which consists noise but they are not like extremely dark and difficult to understand. As shown in Figure 5.5 (b). For such Document image Algorithm-II is proposed. These algorithms are given below.
Algorithm 1 - Steps to remove noise from Type 1 Document image:

1. Convert the document image into gray scale image.
2. Apply Histogram equalization
3. Apply 3X3 Max filter
4. Apply thresholding
5. Apply Frequency domain Gaussian Lowpass filter

Figure 5.5 (a) Type 1 Image; with small dynamic ranged (b) Type II image

Figure 5.6 illustrate the above algorithm for noise removal from dark and noisy image in which (a) shows original document image which is dark, (b) is histogram equalized image of original image. In Figure 5.6 (c) and (d) a 3X3 Median and Max filters of (b) are shown respectively. It can be observed that the 3X3 Max filter gives better results than median filter in case of dark document image. Similarly image (e) shows thresholded binary image of image (d). And finally Frequency domain Gaussian Low-pass filter is applied on (e) which removes the high intensity pixels and blurred the image which produces a noise free document image as sown in (f).
Figure 5.6  (a) Original Image, (b) Histogram equalized image (c) 3X3 Median filter of image b (d) 3X3 Max filter of image b, (e) thresholded image of d (f) Frequency domain Gaussian filter of image e.

Algorithm II- Steps to remove noise from Type 2 Document image:
1. Convert the document image into gray scale.
2. Apply 3X3 Max filter on gray scale image
3. Apply thresholding on result of step (2)
   o If image consist of high intensity foreground pixels then Apply Frequency domain Gaussian Lowpass filter on result of (3).
   o Else go to step 4
4. Obtained Noise free image.
Figure 5.7 (Left) Noisy images (right) results of Algorithm II.

Figure 5.7 illustrate the results of above algorithm in case of the noisy image without very dark background. In Figure 5.7 Left column shows the noisy image and right side column shows the resultant image after applying Algorithm II on noisy images. In proposed method couples of algorithms are implemented for Noisy document image and very dark noisy document image. Spatial and frequency domain filters are used for noise removal. The Max and Median filters of 3X3 neighborhood are used for filtering. Also histogram equalization technique is applied on dark images for normalization purpose. Frequency domain Gaussian Low-pass filter is used for smoothing of the resultant image.

5.1.5 Connecting Broken Characters:

Frequently the characters to be recognized may be broken due to different reasons. The broken characters may produce incorrect recognition and classification of these characters. The broken character may also look similar to other character and there is high probability of recognition as wrong class member due to common features. Connecting broken character is necessary for getting satisfactory results. Broken characters can be connected by applying filter using specific mask. The type of mask depends on the type of broken character or script of broken characters. In current work the broken characters are connected through morphological operations.
and median filters. Here is combination of dilation, erosion, image opening and closing are found useful for connecting broken characters with different structuring elements. Figure 5.8 (a) shows the broken Urdu characters and Figure 5.8 (b) shows the connected character after applying morphological opening using octagon type structuring element.

![Figure 5.8](a) Broken Urdu characters and (b) connected characters after applying morphological opening

5.1.6 Skew Normalization

In preprocessing, the normalization intended to remove numerous variations from the document image, and get standardized image for better feature extraction. The reasons for these variations in document image may be due to improper scanning, skewness, slants, handwriting variations, variations in font type and size etc. These variations directly effects on the accuracy of feature extraction and recognition process. For a robust Character Recognition system such variation must be normalized and standardized using different techniques. A typical normalization module is designed for Skew Normalization, Baseline extraction, Slant Normalization, Size Normalization and contour Smoothing.

*Skew Angle Estimation and correction* is significant process because majority of the character recognition systems are sensitive to skew present in document images. Only small skew angles can be tolerated without any preprocessing. Significantly a skew angle detection and correction is required to improve the results of segmentation and the character recognition. Typically Skew angle is the angle that the text lines in the digital image make with the horizontal direction. Therefore, skew estimation and correction are important steps before line and words segmentation. At the time of scanning the document is fed through scanner, it may cause for skewness of tilted document. The reasons for such tilted image and skewness may be the human error while placing the document page on scanner. Manually we can identify the document skew with certain angle, but in case of slight degree of skew angle, it may
not be possible to detect by human vision. The skew type may vary according to tilted area in the document. If a whole page block is skewed with same orientation then it is called global skew, on the other hand if various blocks have different slant then other block it is called non-uniform text line skew. Finally skew come into view due to fluctuation in line orientation at different locations of line. Most of the system focuses on the global skew angle detection and correction methods [42]. Figure 5.9 shows the skewed handwritten Urdu document Images.

![Figure 5.9 A paragraph of Handwritten Urdu Skewed Document Images.](image)

5.1.6.1 Various De- skewing Techniques:

The line segmentation becomes more complex in case of skewed document image, because the lines are overlapped horizontally and for segmentation blank space or white pixels can’t be found between them. Specifically for projection profile based segmentation, the skewed pages lead to wrong segmentation results. For such segmentation methods it is necessary to remove the skewness from the document image. In literature numerous methods for skew detection and correction are present. Most of the techniques are based on Hough transform, Projection profile, Fourier transform, Cross-correlation, Center of gravity – COG, Nearest neighbor connectivity, Linear regression analysis and Mathematical morphology, Gradient orientation histogram Figure 5.10 shows few of these techniques.

<table>
<thead>
<tr>
<th>Skew Normalization Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Projection Profile</td>
</tr>
<tr>
<td>Fourier Transform</td>
</tr>
<tr>
<td>Nearest Neighbor Connectivity</td>
</tr>
<tr>
<td>Cross-correlation</td>
</tr>
<tr>
<td>Centeroid-COG Based</td>
</tr>
<tr>
<td>Linear Regression Analysis</td>
</tr>
<tr>
<td>Hough Transform</td>
</tr>
<tr>
<td>Mathematical Morphology</td>
</tr>
</tbody>
</table>

![Figure 5.10 Various Skew Normalization Methods](image)

Most of the skew detection methods have the following common features: (1) a prior text/graphics separation is necessary, which may take a significant amount of
Preprocessing and Segmentation

time, though it can be useful for the next steps; (2) large text areas have to be present
on a page for an accurate estimation; (3) many techniques have been designed for
high-resolution images ranging from 100 to 300 dpi [43]. In projection profile
technique a series of projection profiles are obtained at a number of angles close to
the expected orientation, and the variation is calculated for each of the profiles. The
profile that gives maximum variation corresponds to the projection with the best
alignment to the text line; this projection angle is called the skew angle [44]. The
Hough transform [45] is another popular technique for skew detection; this transform
is often applied to a number of representative points of characters such as the
lowermost pixels or centers of gravity. Each representative point (x,y) is mapped from
the Cartesian space to the points (ρ,θ) in the Hough space by forming a set of lines
coming through (x,y) with a slope ρ and distance θ from the origin. The skew
corresponds to the angle associated with a peak in the Hough space. The high
computational complexity of the Hough transform often imposes restrictions on the
possible angle range.

Proposed Skew Normalization Method:

In current research work moment based approach is preferred for skew
detection. An attempt has been made to rotate and de-skew the document image in an
ideal position which was scanned with certain skew angle. Four farthest points have
been calculated using most upper left, upper right, lower left and lower right blank
pixels. Instead of rectangular area polygon area is considered, because handwritten
document always doesn’t consists a rectangular area. Moment based approach has
been used to find centroid i.e. CoG. A baseline has been drawn from CoG to the
origin of the image. Skew angle has been calculated using the slope of the baseline.
The de-skewing method proposed in [46] has been considered as base for our
technique. The text document is inscribed by considering the farthest pixel in the four
directions and determined the rectangle. Text image may contain more than four
corner points, in that case text image has to be considered as polygon. Although a
rectangle is used as a polygon, it can be calculated as:

\[
A = \frac{1}{2} \sum_{i=0}^{n-1} (x_i y_{i+1} - x_{i+1} y_i) 
\]

(8)

Where A is the Area of polygon,
Then we have adopted, moments based method proposed in [47] and applied it for computing skew angle using extracted coordinate. Moments are measures of the pixel distribution around the center of gravity of the document image and allow capturing the global shape information. For discrete 2D image, the moments are evaluated using equation 8. Further, for the binary image, moments of the image function f(x, y) are evaluated in equation 9 as follows:

\[ m_{pq} = \sum_{u} \sum_{v} x^p y^q \]  

(9)

Where for p, q = 0, 1, 2,… and n * n is the size of the image.

From equation 9, central moments for discrete image can be calculated in equation 10, equation 11 for binary image

\[ \mu_{pq} = \sum_{u} \sum_{v} (x - \bar{x})^p (y - \bar{y})^q f(x, y) \]  

(10)

\[ \mu_{pq} = \sum_{u} \sum_{v} (x - \overline{x})^p - (y - \overline{y})^q \]  

(11)

\[ m_{10} = \sum_{u} \sum_{v} x^1 - y^0 \]  

(12)

\[ m_{01} = \sum_{u} \sum_{v} x^0 y^1 \]  

(13)

\[ m_{00} = \sum_{u} \sum_{v} x^0 y^0 \]  

(14)

Using results of equation 12, 13 and 14 we compute equation 15 and 16

\[ \overline{x} = \frac{m_{10}}{m_{00}} \]  

(15)

\[ \overline{y} = \frac{m_{01}}{m_{00}} \]  

(16)

\[ \mu_{11} = \sum_{u} \sum_{v} (x - \overline{x})^2 - (y - \overline{y})^0 \]

\[ \mu_{11} = m_{20} - \frac{2m_{10}^2}{m_{00}} + \frac{m_{10}^{10}}{m_{00}} \]

\[ \mu_{11} = m_{11} - \bar{x} m_{01} \]
The theta (skew angle) is computed using equation 17, 18, and 19 and it results in equation 20.

\[
\theta = \frac{1}{2} \tan^{-1} \left[ \frac{2\mu_{11}}{\mu_{20} - \mu_{02}} \right]
\]  

Equation 20 gives the \( \theta \) for the set of points. The skew angle is estimated using this equation.

**Algorithm:**

The proposed method comprises of subsequent seven steps, where Urdu document image is an input and expected output is de-skewed Urdu document image.

- Four farthest points are determined in possible four directions as shown in Figure 5.11 (b).
- Analyze all farthest points to find whether above two point are present same row or not, Similar process for lower farthest points. In this way it can be observed whether image is skewed or not as shown in Figure 5.11 (a).
- Area of polygon of image is calculated.
- Centroid – Centre of Gravity of image is calculated using values of \( \bar{x}, \bar{y} \) coordinates of all four point with help of equation 15 and 16 respectively Figure 5.11 (c).
- The moments are calculated up to second order.
— Find the angle $\theta$ using equation 20, which is in fact the skew angle of the image Figure 5.11 (c).
— Rotate the document in the clockwise or anticlockwise direction with determined skew angle as shown in Figure 5.11(d).

The algorithm has been tested on a set of 125 different types of Urdu documents images. Various handwritten text document, Urdu news paper, magazine, text books images with skewed angle were tested. The proposed method applied on both handwritten and the printed documents. By analyzing the result in Table 5-2, it is clear that the proposed method work much better on printed documents as compare to handwritten document images. But most of the time inaccurate results are found while detecting whether document is skewed or not i.e. second step of algorithm. The results are reported in Table 5-2 for different types of documents.
Table 5-2 Accuracy percentage with respect to document’s type

<table>
<thead>
<tr>
<th>Document Type</th>
<th>No. of Images</th>
<th>Average Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Handwritten Page</td>
<td>24</td>
<td>75%</td>
</tr>
<tr>
<td>Newspaper</td>
<td>12</td>
<td>80%</td>
</tr>
<tr>
<td>Magazine</td>
<td>10</td>
<td>90%</td>
</tr>
<tr>
<td>Text Book</td>
<td>20</td>
<td>80%</td>
</tr>
</tbody>
</table>

5.1.7 Size Normalization

The accuracy of feature extraction can be improved by normalizing size of the dataset. If all training and testing data size gets standardized then pixel alignment of character will be in similar fashion. As mentioned in [48] ‘Normally, the character image is mapped onto a standard plane (with predefined size) so as to give a representation of fixed dimensionality for classification. The goal for character normalization is to reduce the within-class variation of the shapes of the characters/digits in order to facilitate feature extraction process and also improve their classification accuracy. Basically, there are two different approaches for character normalization: linear methods and nonlinear methods’. If sizes of training testing characters have not normalized before feature extraction then size and pixel alignment of each character will vary, and it may leads into wrong feature extraction recognition. The normalization maximizes the recognition rate in the training data because the large intra-class differences in length of words can be avoided by normalizing training dataset. Figure 5.12 (a) shows sample character ‘Bey’ with variation in size and scale and Figure 5.12 (b) shows normalized data set of these characters.

Figure 5.12 (a) Urdu Character ‘Bey’ with variation in size and scale (b) 60 × 60 size and scale Normalize Urdu Character ‘Bey’
In current research work, aspect ratio adaptive normalization (ARAN) technique for character normalization proposed in [48] has been adopted. The aspect ratio of the normalized character (R2) is adaptively calculated based on that of the original character (R1) using a mapping function. If W1, H1 are the width and height of original characters respectively and width and height of normalized character are w2, H2 respectively then the aspect ratio of original image (R1) and aspect ratio of size normalized image (R2) can be calculated by

\[
R1 = \begin{cases} 
\frac{W1}{H1} & \text{if } (W1 < H1) \\
\frac{H1}{W1} & \text{if } (H1 < W1)
\end{cases} \quad (21)
\]

\[
R2 = \begin{cases} 
\frac{w2}{H2} & \text{if } (W1 < H1) \\
\frac{H2}{w2} & \text{if } (H1 < W1)
\end{cases} \quad (22)
\]

Once the aspect ratio of original character image (R1) is calculated then the aspect ratio of normalized character image can be calculated using various mapping functions for the aspect ratio mapping. The primary strokes and secondary stroke image were normalized into size of 60X60 and 20X20 respectively.
5.2 **Segmentation**

Once the preprocessing is done, the document image turns into noise-free and normalized image. The next step in typical character recognition is segmentation, in which the document image decomposed from page till word and character. Preprocessing facilitate the segmentation by removed skewness, noise and normalizing document. Without which, the document segmentation may leads into wrong direction. Segmentation is a significant process which provides the actual words and characters for extracting features and recognizes them. If segmentation process is not performing properly then it may be garbage input for feature extraction and recognition process. It may be garbage input and garbage output process if wrongly segmented images are provided as an input to feature extraction module. Segmentation approaches are varies intend to each script in character recognition. A segmentation approach suitable for Devanagari script may not be effective for other scripts like Arabic. For complex document two level segmentations are required, External Segmentation and Internal Segmentation.

5.2.1 **External Segmentation:**

If a document structure is complex; it consists of multiple components like, Text, Non-text, Graph, Columns etc. then external segmentation is applied to decompose it. This segmentation decomposes the complex layout of page into logical units. It is a part of document layout analysis; the document layout analysis is a special research domain with its own methodologies and techniques. Before segmenting text into words and character, it is necessary to recognize the text and non-text regions of the document image. Various decomposition methods are proposed by researchers for such decomposition. In external segmentation page layout analysis is perform in to steps structural analysis and functional analysis. In structural analysis decompose the document into paragraph, words, lines and characters. Whereas functional analysis carry out logical labeling like Heading, caption, author name, title, abstract etc. The external segmentation comes under document analysis due which it is not categorized into text segmentation method.
5.2.2 Internal Segmentation:

Unlike external segmentation the internal segmentation deals with decomposition of image consists a sequence of characters into isolated characters or symbols. These segmented characters then forwarded for extracting their features and recognition. The wrong segmentation may lead to inaccuracy in feature extraction and recognition. The fundamental strategies of segmentation can be categorized as following. These strategies may also used in combination, which is known as hybrid approach. [49]

- Dissection / Classical Approach
- Recognition base segmentation
- Mixed Strategy

5.2.2.1 Dissection (Classical Approach):

This methodology is also known as explicit segmentation technique or classical approach. Under this type of segmentation methodology the segments are identified based on character like properties. This process of cutting up the image into meaningful components is given special name “dissection” [50]. In this approach the input image is decomposed into meaningful sub-images based on general features or properties similar to the characters. This process analyzes an image without using specific class of shape information. Usually segmentation principle is agreement of common properties of the segments with those expected for valid characters. General properties which are considered for segmenting can be width, height, disposition along a baseline, and separation from neighboring components etc. which should be satisfactory for valid character features. It is a content dependent method, which uses description of character like shape and contextual knowledge.

This method is more efficient in strictly constrained writing environment. Where writers are restricted to write into specific printed box, or within specific coordinate’s area of the page. To make segmentation easy and simple, numerous dissection base techniques were implemented and can be observed in earlier literature but problem is that; these are strictly constrained writing which doesn’t provide freedom for user. Popular techniques under dissection base method are white space and pitch, vertical projection analysis, connected component analysis landmarks.
5.2.2.1 Fixed pitch

Pitch is the number of characters per unit of horizontal distance. Fixed-pitch font is a font whose letters and characters each occupy the same amount of horizontal space. For printed character recognition system the text is printed with limited font sets and fixed pitch or fixe width which helps to segment the successive characters because each character occupy the same amount of horizontal space. Basically fonts are of two types, fixed width fonts and proportional fonts. In case of fixed-width font all characters have same width, for example character “i” and “m” will be of fixed width and take same horizontal space. Courier New is example of fixed width font. On the other hand in proportional font characters are of dissimilar widths. These fonts are convenient to read and look more professional than fixed width fonts. Figure 5.13 shows fixed width font and proportional font, which illustrate that in fixed-width font every character take fixed horizontal space.

![Fixed width and proportional fonts](image)

*Figure 5.13 Fixed width and proportional fonts [51]*

The characters are printed at approximately equal distance which helps to estimate and find the segmentation point from where the character is to be segmented. For handwritten or hand printed fixed pitch, separate boxes are provided for placing the individual characters. The fixed pitch text segmentation give improved results in many cases where unconstrained text document are partially succeed to segment. In case of poor resolution, broken or merged characters etc. pitch segmentation can be efficient because it cut the character iteratively in a loop by assuming constant width.

5.2.2.1.2 Projection Analysis:

In this segmentation method vertical projection profile of the document image is used. Vertical histogram of the black pixels present in document image taken. Histogram based technique counts the total number of black pixels in each column of the image. Number of present black pixel in each column decides whether it is a
segmentation point or not. Specific threshold value can also be used for segmentation, if total number of black pixel are less than threshold value then it is consider as segmentation point and image is split into sub-image. If the count is greater than predefined threshold value then the segmentation is not performed. In this way the white spaces between successive letters can be found. It is one of the simple and efficient methods but it can be effective in case of non-cursive scripts.

### 5.2.2.1.3 Connected Component Analysis

In unconstrained environment projection profile and fixed pitch based segmentation not efficient. These methods are useful for constrained printed or handwritten text. For such unconstrained environment one of the popular dissection methods is connected component analysis. Basically all character is made-up of black pixels which are connected with each other. In some character which consist secondary strokes, the black pixels of secondary strokes are completely not connected. The pixels are connected with their neighboring pixels in any direction. The connected components may be of 4 connected or 8 connected according to direction. In connected component analysis based segmentation method labels are assigned to every connected group of pixels. And all the connected pixel regions are segmented using pixel and coordinate information. The method can be implemented by using bounding box. The bounding box is refers to the smallest rectangle which entirely encloses a completely connected component. Bounding box can be used to segment non-cursive characters made-up of connected pixels. But in case of characters which consists secondary components it doesn’t work efficiently. The preseason is that, secondary strokes are segmented as separate character, whereas it is the part of another character. Similarly if a character is multi-stroke or broken then bounding box will be drown around each broken part or secondary stork of character. It is also unsuitable in cursive script where characters are touching successive characters then these may be considered as single connected component. And it will be segmented by considering it as single character.

### 5.2.2.1.4 Splitting of connected components

Segmentation of cursive, touching, overlapping characters is more complex as compared to unconnected characters of word is easy to segment. Such touching
characters are segmented as considering a single character due to unavailability of vertical blank pixel. For this type of scripts the unique features of such intersecting characters are observed and stored in knowledge base. By detail analysis of image and Information from taken from knowledge base used to split these two or more intersecting or overlapping characters. In this method details examination of connected component i.e. bounding box is carried out which may believe an extension of bounding box analysis. Knowledgebase play an important role to decide whether connected components belongs to Digits, capital letters, zip code, mobile number etc and help to segment is appropriately. For example account number can be easily recognized as it consist specific range of characters; similarly how many digits are there in any mobile number. This prior information may be useful for segmenting or splitting connected components which are basically individual symbols but connected due to some reasons. This approach analyses connected components using simple recognition logic and detect the connected components which are probably broken, connected or single character components. Also a pre-recognition modules are applied recognized whether it is connected characters or not, then it is segmented. Afterwards proper splitting path is determined and followed to segment the touching characters.

5.2.2.2 Recognition base segmentation:

The recognition-based method is an implicit technique, where the characters are segmented at the same time as being recognized, due to which it is known as straight segmentation or recognition based segmentation. The specific rules are used to spot segmentation point of every character. One of the major benefits of this method is that it avoids serious classical character separation and no exact character segmentation path is essential. Due to which it is suitable for character recognition of scripts which has serious segmentation problem. This method generally uses a dynamic width mobile window which provides a sequence of tentative segmentations which are then confirmed (or not) by the character recognition as a result of a coherent segmentation/classification result [52]. This method is also known as segmentation-free approach [53]. Cheung et al. [54] presented this type of recognition-based segmentation algorithm in which the word bitmap is processed sequentially in a step-by-step mode. At every step character is ensure for recognition against a pre-specified feature space.
Once it is being recognized, the character is then isolated. This method has a disadvantage of involving high computations requirement with accordingly low recognition speed [55]. In Recognition based segmentation designing and implementation of dissection algorithms is not required and complexity may reduces. The classifiers are used directly in this method, where a dynamic width mobile window blindly divides input image into numerous overlapping pieces and chooses the proper segmentation upon the recognition confidence. Initially the windowing is applied on complete document image and segmentation hypotheses are generated. After that in verification, the best hypothesis is chosen which is determined by classifier.

Kurt Alfred Klueve [56] discussed a range of approaches for Recognition-based segmentation in which it is mentioned that, in recognition based segmentation first windowing is performed on the image to generate segmentation hypotheses. Later in verification step the best hypothesis given by the classifier is preferred. Early methods used windowing techniques that classified the character based on a prototype character (templating). The system would exhaustively search all possible cut points in the image until all characters had been matched against a prototype library within a given threshold. Further he discussed an approach that uses a shortest path to find the most accurate segmentation of the word. It calculated all the possible segmentation boundaries and constructed a graph, where all paths through the graph resulted in potential matches. The left to right path with the lowest weight corresponded to the best recognition rates.

5.2.2.3 Holistic Approach

Mainly the recognition approaches can be categorized into segmentation free approach (holistic) or segmentation-based. The segmentation based approach encounters several difficulties. To avoid such segmentation complexities, researchers frequently proposed an approach called holistic approach or global approach. In segmentation-based approaches, the characters are recognized individually after they have been extracted from the text image. Unlike the holistic approach uses the top-down approaches to recognize whole word. In holistic approach characters are not segmented or identified individually, instead of character, whole word is considered as single unit and recognized by applying global recognition method. Due to which, it is also known as segmentation-free approach.
Basically this approach was introduced for speech recognition [52, 57]. It completely bypasses the complex dilemma of word segmentation into characters. As in this approach the words are taken as whole instead of segmenting it into strokes or characters, it is called Holistic approach. The holistic analysis is adequate for problems with medium and small vocabulary such as literal amount reading in bank checks [58,59,60]. In this approach paragraphs of document are first splits into separate lines, and then lines are split into words. Usually horizontal projection profile technique is used to split paragraph into lines. Similarly vertical projection profile can be applied to segment lines into words. [61]. Then feature extraction is done to create a representation of the word like directional strokes, diacritical marks, cusps, ascenders, descenders, etc. After which, global recognition is carried out by matching the representation of the word with a representation of a predefined word in the lexicon.

The representation may be feature vector, template of the word etc[56]. Holistic approach is limited for small vocabulary applications and unable to deal appropriately with a large number of classes. Therefore, holistic classifiers can be suitable if it is used in conjunction with other approach. The computational requirement may decrease due to limited vocabulary but due to complexity in representation of whole word the accuracy of reorganization may reduce [50]. Although it is attractive, due to simplicity, this method has few drawbacks. In practical applications, it is not possible to have enough sample data for robust modeling, which typically prevents its application to real-world recognition tasks. Also unable to recognized new words, which are not present in system vocabulary [62]

5.2.3 Urdu Script: Tested Segmentation methods

During research work different segmentation methods are tested on Urdu script. As mentioned earlier that, the segmentation of Urdu script is one of the challenging task in character recognition. It is found during literature survey that most of the researchers [63] preferred holistic approach instead of segmentation based approach. But Segmentation free method has its advantages and disadvantages, like it can be useful for limited vocabulary system. As Urdu consists a huge vocabulary and with every passing day new terminologies are being adopted by Urdu, that’s why
holistic approach may not be ideal. In this research work an attempt is made to test various segmentation methods on Urdu script to find best one. Various methods like, Projection profile, baseline extraction, connected component analysis have been tested. But every method has limitation due to cursive nature and word as well as line overlapping in Urdu text. An attempt is made to find the reasons behind failure of individual segmentation methods in case of Urdu text. The techniques which have been tested are

- Projection profile
- Baseline Extraction
- Connected Component Analysis

5.2.3.1 Urdu Script and Projection Profile based segmentation

Projection analysis is one of the simple and efficient methods but it can be effective in case of non-cursive scripts. In case of non-cursive printed text, the projection profile method can be one of the best segmentation methods. As shown in Figure 5.14 all the characters in printed English text can be easily segment using projection profile method. But Urdu is a cursive script in which projection analysis method is not satisfactory, the reasons are horizontal overlapping and touching characters. Figure 5.15 shows the failure of projection profile method in case of cursive Urdu text. The accuracy of projection profile based segmentation is poorer in case of handwritten Urdu text as shown in Figure 5.16.

Simple vertical projection analysis the histogram based method doesn’t work effectively for tilted text like Urdu. Another limitation is that, the estimation of minimal local becomes extremely complicated in case of overlapping or touching Urdu characters. One more complication in Urdu character segmentation is the presence of diacritic and secondary components, which may generate false segmentation points. Vertical projection analysis method is not much suitable in case of handwritten cursive text because it consists of slants, overlapped, broken and touching characters. For such type of complex text, various methods were proposed by researcher to resolve this dilemma. Few of them are directional projection, pre-filtering, and fixed thresholding, peak-to-valley function, removing secondary components before taking projection profile, Maxima and minima analyses of the vertical projection histogram. In few methods instead of entire word only vertical
projection profile of specific zones is considered. Like upper, middle and lower zones are considered. However in our experimental work these methods are only partially successful for Urdu script. Segmentation due to overlapping, secondary strokes, diacritic and absence of accurate baseline are the primary reasons behind slow progress in Urdu character recognition system.

5.2.3.2 Problem of baseline extraction in Urdu

A baseline is a virtual line over which the writer writes the word. The baseline is primarily a virtual line on which all the characters of cursive or semi cursive writing are aligned or joined. A baseline can also be defined as ‘a line where most of the pixels in the word/text line are concentrated’. In case of Devanagari scrip
shirvrekha is like a base line where all characters of a word are joined. For scripts like Arabic the modified forms of alphabets join on the baseline. In Arabic character recognitions many researchers used baseline for slant correction and segmentation. Arabic words are aligned or joined on an imaginary line is known as baseline. In Arabic script baseline is frequently used by many researchers to de-skew line, segmentation etc.

![Figure 5.17 Arabic language character shapes based on baseline [64]](image)

**Figure 5.17 Arabic language character shapes based on baseline [64]**

(a)  
(b)  
(c)  

![Figure 5.18 Baseline extracted using projection of (a) Printed Arabic (b)Printed Urdu and (c) Handwritten Urdu – poor baseline](image)
Atallah et. al [64] presented a Comparative Study between various Arabic Baseline Detection methods, in which it is cited that using baseline, the characters and shape are classified into three groups Ascenders, descendiers and special marks called diacritics such as dots, shadda (Zigzag) and maddah, and these groups may be constructed from stroke or small element or complete character. The Arabic language character shapes based on baseline shown in Figure 5.17. Ascenders lie above the baseline, but descendiers lie under the baseline, and special marks lie in either above or under the baseline depending on the character [65,66]. But in case of Urdu script absence of a reliable baseline is the big problem. Venu Govindaraju et. al. [67] mentioned that, Urdu Connected characters do not follow a baseline because the joined characters in Urdu are positioned according to their preceding, succeeding, and vertical justification of the ligature. As shown in Figure 5.18 (a) the baseline in printed Arabic text can be calculated using projection profile method, because most of the words are aligned on single straight line. But in case of Urdu text, extracting reliable base line using projection profile doesn’t provide satisfactory results.

Comparatively base line detection in printed Urdu text may give little satisfactory results as illustrated in Figure 5.18 (b). But Detection of baseline is more complex in case of handwritten Urdu text, because all the characters and words of a text line don’t aligned or join at a single base line. As shown in Figure 5.18 (c) extraction of baseline for handwritten Urdu text, using horizontal projection profile gives poor results. Extraction of such poor base line affects the segmentation process in Urdu handwritten text, which makes Urdu character recognition more complex and challenging.

5.2.3.3 Urdu Script and Connected Component Analysis:

The method is implemented by using bounding box. The bounding box refers to the smallest rectangle which entirely encloses a completely connected component. Bounding box can be useful in non-cursive script characters without secondary strokes, which is made-up of connected pixels. In case of Urdu script any unbroken, single stroke and isolated character like ۡ, ١, ٢ etc. which doesn’t have secondary components can be easily segmented using this technique.

But in case of characters which consists secondary strokes it don’t work efficiently because secondary strokes are segmented as separate character, whereas it
is the part of another character. Like in case of character (ٍ ، ٍ، ﯦ، ﯦ) , these will be segmented into two or more parts, the primary stork and other will be secondary stork i.e. upper dot, lower dots, diacritical marks etc. Figure 5.19 (a) shows a single stroke isolated Urdu characters, which are easily segmented through connected component analysis. But this segmentation approach is unsuitable for multi-stroke or multi-component characters as shown in Figure 5.19 (b). Where primary strokes are segmented as separate character and also enclosing secondary strokes as separate characters i.e. wrongly segmented.

If a character is multi-stroke or broken, then bounding box will be drowning around each broken part or secondary stork of the character. Similarly connected component analysis is not suitable to segment Urdu word, line and paragraph segmentation as shown in Figure 5.20 a, b and c respectively.

![Figure 5.19 Connected Component Analysis based segmentation - (a) Single stroke Isolated characters, (b) Multi-stroke characters](image)

This method is also inadequate because in cursive script if character is touching successive characters then these may be considered as single connected component. And it will be segmented by considering it as single character. There are various reasons due to which a simple connected components analysis using bounding box approach may not be an efficient approach few of them are cursive Urdu script, multi-stroke characters, and broken characters. Due to which, in current research work it is used with combination of other technique to get satisfactory results in Urdu script text segmentation.
Reference:


20. Mandana Kavianifar, Adnan Amin, “Preprocessing and Structural Feature Extraction for a Multi-Fonts Arabic / Persian OCR”, IJDR 1999, PP. 213-216


60. Hassiba Nemmour and Youcef Chibani, “Handwritten Arabic Word Recognition based on Ridgelet Transform and Support Vector Machines”.


