CHAPTER 4

DETECTION AND LOCALIZATION OF TEXT FROM HETEROGENEOUS TEXTUAL IMAGES

4.1 INTRODUCTION

The extraction of a text may seem to be a trivial application for existing optical character recognition (OCR) tools. However, compared to OCR for document images, extracting a text from real images faces numerous challenges, due to lower resolution, unknown text color, size and position, or complex backgrounds. Text extraction and recognition, which include text detection, localization, segmentation and binarization, and recognition, is a useful process for text-based image indexing. Furthermore, it is a very important task in searching for information in web sites or digital multimedia libraries (e.g. databases of images, videos or document images). The first three processing stages are important to achieve high-quality text recognition results when applying an OCR system. This chapter is focused on the first two problems of TIE system namely text detection and localization.

Text extraction from Images starts with detecting and locating the text in an image. The common issue is the problem of finding a decision function for each pixel, and deciding whether it is part of a text or not. This chapter considers the problem of identifying text regions in images. It is observed from the literature that
• The existing Text detection / localization (TD/TL) system enforces a separate methodology to be devised for scene text images, caption text images and document images. A single methodology handling all three types of images has not been reported so far.

• No algorithm detects the text properly from images with angular text, perspective projection / transformation and radially changing text

• Many algorithms make a priori assumptions about the text to be extracted (e.g. strong restrictions on text color, size, location, etc.) and are not robust for complex backgrounds and variations in lighting conditions.

All the above mentioned variations in a text and the lack of a unified and domain independent text extraction system for heterogeneous textual images demand the development of new methods to address these issues. With this in view, a unified SBTA-TD/TL system introduced in this chapter is focused to automatically detect and localize the text appearing in heterogeneous textual images such as Scene text images, Caption text images and document images with a common framework, and generate a bounding box around a localized text which will in turn be given as input to the Text segmentation and Binarization methods.

4.2 ALGORITHM A: SUB BAND TEXTURE ANALYSIS BASED TEXT DETECTION / LOCALIZATION (SBTA – TD/TL)

The morphological approach for text detection and localization proposed in this chapter, works across different kinds of images with a unified framework and takes care of limited font sizes and orientation of text, and eliminates the need to devise a separate methodology for various kinds of
images. Here, the idea is to apply a non sub sampled contourlet transform NSCT (Arthur et al 2006) on images to capture Multi–oriented texture details, instead of recognizing the edges in only horizontal, vertical and diagonal directions. These details are captured at high frequency component so as to produce text regions whereas at low frequency gives rise to non-text region. This thesis has been motivated by the fact that the texture details in the decomposed sub bands can be analyzed and used as the distinguishing factor to separate text and non text. The objective of this thesis is to design Image analysis based SBTA-TD/TL technique, which eliminates the process of connected component analysis thereby decreasing computational complexity (pixel access). It adapts to various heterogeneous textual images due to its Multi oriented texture details.

4.2.1 Image Decomposition and Texture Analysis

Texture-based methods examine the local texture features within small regions of an image. The text present in images exhibits some distinct textural properties, which may be used to distinguish it from the background. Gabor filters, Wavelets, Fast Fourier transformation, etc. are usually used to extract the textural properties of a text region in an image. If the texture features are consistent with the characteristics of a text, all the pixels in the region are marked as text. Image decomposition is useful for a host of image processing tasks, e.g. texture segmentation and image inpainting. The various transforms used for image decomposition are briefly described here.

4.2.1.1 Fourier Transform

The Fourier Transform is an important image processing tool which is used to decompose an image into its sine and cosine components. The output of the transformation represents the image in the Fourier or frequency domain, while the input image is the spatial domain equivalent.
The DFT (Discrete Fourier Transform) is the sampled Fourier Transform, and therefore, does not contain all the frequencies forming an image, but only a set of samples which is large enough to fully describe the spatial domain image. Another sinusoidal transform (i.e. transform with sinusoidal base functions) related to the DFT is the Discrete Cosine Transform (DCT). The main advantages of the DCT are that it yields a real valued output image and that it is a fast transform. The Fourier analysis has a serious drawback. When a signal is transformed into the frequency domain, time information is lost. The Short-Time Fourier Transform (STFT) maps a signal into a 2-D function of time and frequency. However, the time and frequency information can only be obtained with limited precision. The precision is determined by the size of the window used to analyze the signal.

4.2.1.2 Wavelet transform

The Wavelet analysis is a windowing technique, similar to the STFT, with variable-sized windows. It allows the use of long time intervals when more low frequency information is sought, and shorter regions when more high frequency information is what you are after. The Wavelet analysis is capable of revealing aspects of data that other signal analysis techniques miss, including aspects such as trends, breakdown points, discontinuities, and self-similarity. It is also often used to compress or denoise a signal without any appreciable degradation.

Wavelet transforms have advantages over traditional Fourier transforms for representing functions that have discontinuities and sharp peaks, and for accurately deconstructing and reconstructing finite, non-periodic and/or non-stationary signals. Wavelet transforms do not have a single set of basis functions like the Fourier transform, which utilizes just the sine and cosine functions. Instead, wavelet transforms have an infinite set of possible basis functions. Thus the wavelet analysis provides immediate access
to information that can be obscured by other time-frequency methods such as the Fourier analysis. Wavelet transforms are classified into discrete wavelet transforms (DWTs) and continuous wavelet transforms (CWTs). They can be used to represent continuous-time (analog) signals. CWTs operate over every possible scale and translation, whereas DWTs use a specific subset of scale and translation values or representation grid.

4.2.1.3 Contourlet Transform

Although the wavelet transform is powerful in representing images containing smooth areas separated with edges, it cannot perform well when the edges are smooth curves. New developments in directional transforms, known as contourlets in two dimensions, which have the property of capturing contours and fine details in images can address this issue. The contourlet transform is an extension of the wavelet transform, which uses multi scale and directional filter banks. Here, images are oriented at various directions in multiple scales, with flexible aspect ratios. The contourlet transform effectively captures smooth contour images that are the dominant feature in natural images. The main difference between contourlets and other multi scale directional systems is that the contourlet transform allows for different and flexible number of directions at each scale, while achieving nearly critical sampling.

In addition, the contourlet transform uses iterated filter banks, which makes it computationally efficient; specifically, it requires $O(N)$ operations for an N-pixel image. The contourlet transform (Do and Vetterli 2005) is a multidirectional and multi scale transform that is constructed by combining the Laplacian pyramid (Burt and Adelson 1983, Do and Vetterli 2003) with the directional filter bank (DFB) proposed in Bamberger and Smith (1992).
4.2.1.4 **Non sub sampled Contourlet Transform (NSCT)**

Due to the down and up samplers present in both the Laplacian pyramid and the DFB, the contourlet transform is not shift-invariant. An over-complete transform the non sub sampled contourlet transform (NSCT) has been proposed in the literature (Arthur et al 2006) and applied in the proposed system. The NSCT is a fully shift-invariant, multi scale, and multi direction expansion that has a fast implementation. Here, filters are designed with better frequency selectivity, thereby achieving better sub band decomposition.

4.2.2 **Observations from Various Transforms**

From the above discussion, the characteristics of various transforms are observed and summarized as follows:-

- **Fast Fourier transform (FFT)**
  - When a signal is transformed into the frequency domain, time information is lost.
  - Can not represent functions that have discontinuities and sharp peaks.

- **Short-Time Fourier Transform (STFT)** maps a signal into a 2-D function of time and frequency. However, the time and frequency information can only be obtained with limited precision.

- **Wavelet analysis**:
  - capable of revealing aspects of data such as trends, breakdown points, discontinuities, and self- similarity
  - Used to compress or denoise a signal without any appreciable degradation.
- Cannot perform well when the edges are smooth curves.

In this work, it is decided to apply the texture-based method to have a unified and common framework for all kinds of images. To achieve this, variations in the text of a scene text image are to be handled in a good manner. One main aspect of the variation in a scene text is the variability of the orientation of the text compared to caption text image. To design a common framework for text extraction from heterogeneous images, an appropriate transform should be chosen to decompose the images which can do the following:

- Represent functions that have discontinuities and sharp peaks, and accurately deconstruct and reconstruct finite, non-periodic and/or non-stationary signals.
- Can be powerful in representing images with smooth edges/curves.
- Capture contours and fine details in images.
- Capture multi-oriented edge details on various scales.

From this point of view, the literature is further explored and a suitable transform for the above requirements has been found as a variation of the contourlet transform, namely, the Non sub sampled contourlet transform (NSCT) as follows:

- Contourlet transform:
  - images are oriented in various directions on multiple scales, with flexible aspect ratios
  - captures smooth contours in images
  - is not shift-invariant.
• Non sub sampled contourlet transform, NSCT:
  - a fully shift-invariant, multi scale, and multi direction expansion that has a fast implementation
  - Better frequency selectivity, thereby achieving better sub band decomposition.

Here, the proposed method is to apply the non sub sampled Contourlet transform as a texture based method on images for the first time in literature for the text extraction application. It decomposes the image into a set of directional sub bands with texture details captured in different orientations at various scales. The texture details present in the sub bands are analyzed to detect the text regions by the proposed SBTA algorithm. The proposed system is designed to extract the text from all the three kinds of images and take care of various font sizes and orientation of text, and to eliminate the need to devise a separate methodology for various kinds of images.

The major contributions of the SBTA –TD/TL system are as follows:

• It works across different kinds of images such as Caption text images, Scene text images and Document images.

• It works independently with a limited range in sizes of characters and orientation of text string.

• This can be easily extended to other languages.
4.3 SYSTEM DESCRIPTION

In this section, the processing steps of the proposed SBTA-TD / TL approach are presented. Our aim is to build automatic text region localization and extraction system which is able to accept different types of images such as Caption text, Scene text and Document images. The underlying principle of our method is based on capturing Multi–oriented Texture details at high frequency component so as to produce text regions, whereas at low frequency gives rise to non-text regions.

The Non sub sampled contourlet transform (NSCT) is applied to an input image. It produces $2^n$ sub bands for n level which is specified. Energy is computed for the sub bands, which are categorized into Strong and Weak bands. The boosting level is applied to the weak sub bands so as to bring it to the level of the strong bands. Then, edge detection followed by suitable dilation operation is applied. Strong and boosted edges after dilation are combined with addition which is followed by logical AND operation which forms the text region. Finally, the remaining non text regions are identified and eliminated. Then, binarization is applied to extract the text from the identified text regions. The block diagram of the proposed methodology is shown in Figure 4.1.

![Figure 4.1 Block diagram of the SBTA-TD/TL method](image-url)
The proposed method consists of the following stages: Candidate text region detection, Text region localization and Extraction.

4.3.1 Candidate Text Region Detection

4.3.1.1 Non sub sampled Contourlet Transform (NSCT)

Due to the down and up samplers present in both the Laplacian pyramid and the DFB, the contourlet transform is not shift-invariant. An over-complete transform, the non sub sampled contourlet transform (NSCT), has been proposed (Arthur et al 2006) and applied in our proposed system. The NSCT is a fully shift-invariant, multi scale, and multi direction expansion that has a fast implementation. Here filters are designed with better frequency selectivity, thereby achieving better sub band decomposition.

Figure 4.2(a) displays an overview of the NSCT (Arthur et al 2006). The structure consists of a bank of filters that splits the 2-D frequency plane in the sub bands illustrated in Figure 4.2(b). This transform can thus be divided into two shift-invariant parts: a non sub sampled pyramid structure that ensures the multi scale property and a non sub sampled DFB structure that gives directionality. 1) Non sub sampled Pyramid (NSP): The multi scale property of the NSCT is obtained from a shift-invariant filtering structure that achieves sub band decomposition similar to that of the Laplacian pyramid. This is achieved by using two-channel non sub sampled 2-D filter banks. Figures 4.3(a) and 4.3(b) illustrate the non sub sampled pyramid (NSP) decomposition with J =3 stages. The ideal pass band support of the low-pass filter at the jth stage is the region \([-(([-\pi /2^j], (\pi /2^j)]^2. Accordingly, the ideal support of the equivalent high-pass filter is the complement of the low-pass, i.e., the region \([-((-\pi /2^{j+1}), (\pi /2^{j+1})^2 \setminus (([-\pi /2^j], (\pi /2^j)]^2. The filters for the subsequent stages are obtained by up sampling the filters of the first stage. This gives the multi scale property without the need for an additional filter design. This structure is thus different from the separable non sub sampled
wavelet transform (NSWT). In particular, one band pass image is produced at each stage resulting in J+1 redundancy. By contrast, the NSWT produces three directional images at each stage, resulting in 3J+1 redundancy.

Figure 4.2  Structure of Non sub sampled contourlet transform
(a) NSFB structure that implements the NSCT (b) Idealized frequency partitioning

Figure 4.3  Non sub sampled pyramid decomposition
(a) Three-stage pyramid decomposition (b) Sub bands on the 2-D frequency plane.
Figure 4.4 Four-channel non sub sampled directional filter bank constructed with two-channel fan filter banks  
(a) Filtering structure (b) Frequency decomposition

The filters for the subsequent stages are obtained by up sampling the filters of the first stage. This gives the multi scale property without the need for an additional filter design. This structure is thus different from the separable non sub sampled wavelet transform (NSWT). In particular, one band pass image is produced at each stage resulting in J+1 redundancy. By contrast, the NSWT produces three directional images at each stage, resulting in 3J+1 redundancy.

2) Non sub sampled Directional Filter Bank (NSDFB): The directional filter bank (Bamberger and Smith 1992) is constructed by combining critically-sampled two-channel fan filter banks and resampling operations. The result is a tree-structured filter bank that splits the 2-D frequency plane into directional wedges. A shift-invariant directional expansion is obtained with a non sub sampled DFB (NSDFB). The NSDFB is constructed by eliminating the down and up samplers in the DFB. This is done by switching off the down / up samplers in each two-channel filter bank in the DFB tree structure and up sampling the filters accordingly. This results in a tree composed of two-channel NSFBs as shown in Figure 4.4(a); Figure 4.4(b) illustrates the four channel decomposition. The synthesis filter bank is obtained similarly. The NSCT is flexible in that it allows any number of directions on each scale. In particular, it can satisfy the anisotropic scaling
law. This property is ensured by doubling the number of directions in the NSDFB expansion on every other scale. The NSCT (Arthur et al 2006) is constructed by combining the NSP and the NSDFB as shown in Figure 4.2(a). Here, the image is decomposed by the NSCT and eight sub bands are produced; some of the Non sub sampled contourlet coefficients are shown in Figure 4.5 for Scene text image.

Figure 4.5 Non sub sampled Contourlet coefficients
4.3.1.2 SBTA algorithm

The motivation for the SBTA technique is that the texture details in the decomposed sub bands can be used as the distinguishing factor to separate the text and the non text. This thesis aims to design an image analysis based SBTA-TD/TL technique, which eliminates the process of the connected component analysis, thereby decreasing computational complexity (pixel access) and adapts to various heterogeneous textual images due to its multi oriented texture details.

The original image is decomposed into eight directional sub band outputs using the NSCT at three different scales and the energy of each sub band can be obtained from the decomposed image. The texture details present in the sub bands are analyzed to detect the text regions. The energy of the image block associated with the sub band is defined as

\[ E = \sum_{x=1}^{n} \sum_{y=1}^{m} |I(x, y)|^2 \] (4.1)

Here \( I(x, y) \) denotes the image intensity corresponding to the sub band. The normalized energy value is used instead of the energy value, to avoid threshold inaccuracies due to spatial intensity variations across the image. The sub bands are categorized as strong and weak, based on the value of the Computed Energy. To extract dominant directional energy, it is necessary to select a threshold. Now, the SBTA-TD/TL algorithm determines the difference between each sub band with the maximum energy band as in Equations (4.2) and (4.3).

\[ M = \text{Max}(E_i) \] (4.2)

\[ i = 1, 2, \ldots, 2^j \text{ where } j = 1, 2, \ldots, \text{nlevels} \]

\[ d_{ij} = M - E_i \] (4.3)
Sub bands are arranged in the ascending order based on the difference in the energy in a list, as in Equation (4.4).

\[ l_j = (d_{k1}, d_{k2}, ...., d_{kN}), \]  
(4.4)

where \( d_{k1} < d_{k2} < ... < d_{kN} \).

i.e.) \( set \ l_j = d_{kj}, \) where \( j = 1, 2, ... T ... N \)

The sub bands occupying the first few places from the beginning of the list will have the minimum difference, and will become candidates for strong bands. Then a proper threshold is applied to separate this sorted list into two sets as Strong \((S_p)\) and Weak \((W_l)\) sub bands, as in Equation (4.5) and (4.6). This threshold is determined from the calculated average of the list.

\[ SetThreshold \ T, \]

\[ as \quad T = \text{Avg} (l_j) \]

\[ Wi = l_1, l_2, .......... l_{T-1} \]  
(4.5)

\[ S_p = l_{T+1}, l_{T+2}, .......... l_N \]  
(4.6)

The Energy levels of the identified weak sub bands are boosted with an appropriate boosting factor \((\gamma)\), so as to bring out the proper edges in the edge detection stage as in Equation (4.8). The optimal value of \( \gamma \) is chosen, based on experimentation, and found that it is the ratio of the minimum of the strong sub bands to the average of the weak sub bands, as in Equation (4.7).

\[ AW = \text{Avg} (W_l) \]

\[ Boosting \ factor = \frac{\text{Min}(S_p)}{AW} \]  
(4.7)
\[ W_i \rightarrow \text{Boosted to} \rightarrow \frac{\text{Min}(S_p)}{AW} \]

as \[ W_B = W_i \times \frac{\text{Min}(S_p)}{AW} \]  

(4.8)

where \[ p = T, T + 1, \ldots, N \]

\[ i = 1, 2, \ldots, T - 1 \]

The above procedure for candidate text region detection is summarized as follows:

Declare

\[ k \] : no. of levels

\[ E_i \] : Energy of ith decomposed sub band in image I

\[ N \] : Sub bands count

NSSC : Non sub sampled coefficient

\[ th \] : Threshold

\[ d_{kj} \] : Difference between the energy of jth sub band and the maximum energy

\[ W_i \] : Weak sub bands

\[ S_p \] : Strong sub bands

\[ W_B \] : Boosted weak sub bands

\[ M_S \] : Scaling factor

I. Procedure Candidate text region detection ( )

Step 1. Start

Step 2. For each \( i = 1 \) to \( 2^k \) of I

Compute NSSC coefficients
Step 3. For each \( x = 1 \) to \( n \) of \( I \)

Step 4. For each \( y = 1 \) to \( m \) of \( I \)

\[ E_i = I_i (x, y)^2 \]

Step 5. Next

Step 6. Next

Step 7. Next

Step 8. \( M = \text{Max} (E_i) \)

Step 9. For each \( i = 1 \)

\[ d_{kj} = M - E_i \]

Step 10. Next

Step 11. \( l_j = \text{Sort ascend} (d_{kj}) \)

Step 12. \( th = \text{Avg} (l_j) \)

Step 12. For each \( j \) of \( l_j \)

if \( l_j < th \)

\[ W_i = l_j \]

if \( l_j > th \)

\[ S_P = l_j \]

Step 13. Next

Step 14. \( M_S = \text{Min} (S_P) / \text{Avg} (W_i) \)

Step 15. \( W_B = W_i * M_S \)

Step 16. End

4.3.1.3 Edge detection

The Sobel operator is used to detect the edges for the multidirectional sub bands. The Sobel operator performs a 2-D spatial gradient measurement on an image. Typically, it is used to find the
approximate absolute gradient magnitude at each point in an input grayscale image. The result, therefore, shows how "abruptly" or "smoothly" the image changes at that point, and therefore, how likely it is that that part of the image represents an edge, as well as how that edge is likely to be oriented. The operator uses two 3×3 kernels which are convolved with the original image to calculate the approximations of the derivatives - one for the horizontal changes, and one for the vertical ones. If we define \( A \) as the source image, and \( G_x \) and \( G_y \) as the two images which at each point contain the horizontal and vertical derivative approximations, the computations are as follows:

\[
G_x = \begin{bmatrix}
+1 & 0 & -1 \\
+2 & 0 & -2 \\
+1 & 0 & -1 \\
\end{bmatrix} \ast A \quad \text{and} \quad G_y = \begin{bmatrix}
+1 & +2 & +1 \\
0 & 0 & 0 \\
-1 & -2 & -1 \\
\end{bmatrix} \ast A
\]

where \( \ast \) here denotes the 2-dimensional convolution operation. The \( x \)-coordinate is here defined as increasing in the "right"-direction, and the \( y \)-coordinate is defined as increasing in the "down"-direction. At each point in the image, the resulting gradient approximations can be combined to give the gradient magnitude, using:

\[
G = \sqrt{G_x^2 + G_y^2}
\]

An approximate magnitude can be calculated using: \( |G| = |G_x| + |G_y| \).

Edges have been detected for the multidirectional sub bands and some of the edges are shown in Figure 4.6.

Figure 4.6 Non sub sampled Contourlet edges
4.3.2 Text Region Localization

4.3.2.1 Morphological Dilation

Here, the detected edges of the strong and boosted weak sub bands are dilated. The basic effect of the dilation operator on a binary image is to gradually enlarge the boundaries of the regions of the foreground pixels (i.e. white pixels, typically), by adding pixels to the boundaries of the objects in an image. Thus, the areas of foreground pixels grow in size, while the holes within those regions become smaller. The dilation operator takes two pieces of data as inputs. The first is the image which is to be dilated. The second is a (usually small) set of coordinate points known as a structuring element (also known as a kernel). It is this structuring element that determines the precise effect of the dilation on the input image. Various structuring elements have been experimented with to find the most suitable one and the dilated images are produced using a disk shaped structuring element. Here, dilation is performed to enlarge or group the identified text regions.

4.3.3 Text Region Extraction

The strong and boosted edges after dilation are combined with addition followed by the logical AND operation, which forms the text region as in Equation 4.10. The results of dilation and logical operation are shown in Figure 4.7 b- d. Then, the result is mapped to the original image to get the text regions as shown in Figure 4.7e.

\[ \text{Region} = (W_b \cup S_r) \cap \text{IP} \]  \hspace{1cm} (4.9)

Now, the remaining non text regions are identified and eliminated by removing from a binary image all connected components (objects) that have fewer than \( P \) pixels, producing another binary image, BW2 with CC based thresholding as shown in Figure 4.7f. The default connectivity is 8 for two dimensions \( \text{BW2} = \text{bwareaopen (BW, P)} \). The basic steps are,
1. Determine the connected components.
2. Compute the area of each component.
3. Remove small objects. (CCs having fewer than P pixels)

![Figure 4.7](image)

**Figure 4.7 Steps in text Region Extraction of SBTA –TD/TL method**

(a) Scene text (b) and (c) Dilation after boosting Sub bands (d) AND operation with added result (e) Mapped to Original image. (f) After Removal of non text region.

### 4.3.4 Text Extraction

Binarization is applied to extract the text from the identified text region. It will enable the extracted text to be parsed and recognized by the Common OCR systems. Binarization is a technique by which gray scale images are converted to binary images. The most common method is to select a proper threshold for the image and then convert all the intensity values above the threshold intensity to one intensity value representing either a “black” or “white” value. All intensity values below a threshold are converted to one intensity level and intensities higher than this threshold are converted to the other chosen intensity. This binarization technique segments an image into the foreground and the background. The well known global thresholding method described by Otsu (1979) is tried for binarization. The
foreground contains interested characters and this process generates an output image with a white text against a black background as shown in Figures 4.11 and 4.12. A robust text binarization system towards complex background and variability in text color has been proposed in Chapter 6.

4.4 RESULTS AND PERFORMANCE ANALYSIS

4.4.1 Experimentation Setup

This algorithm has been tested over a corpus of 300 images with 3 sets of images using the 1.79 GHz system having 512MB RAM and various metrics have been evaluated from the tested results. These data set images have been gathered from the sites of several research groups such as Laboratory for Language and Media Processing (LAMP), Automatic Movie Content Analysis (MoCA) Project, Computer Vision Lab., and Pennsylvania State University. The parameters which are considered for experimentation are listed in Table 4.1.

- Set1 contains artificial/Caption text images
- Set2 contains scene text images
- Set3 contains document images

<table>
<thead>
<tr>
<th>Table 4.1 Attributes considered for the Experimentation setup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total no of Images tested</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Type of images taken</td>
</tr>
<tr>
<td>Font sizes tested</td>
</tr>
<tr>
<td>Orientation of text</td>
</tr>
<tr>
<td>Languages of text checked</td>
</tr>
</tbody>
</table>
4.4.2 Performance Analysis

The metrics used to evaluate the performance of the system are Precision, Recall and F-Score. Precision and Recall rates have been computed, based on the number of correctly detected characters in an image, in order to evaluate the efficiency and robustness of the algorithm. The metrics are as follows:

Definition 1: The Precision rate (P) is defined as the ratio of correctly detected characters to the sum of correctly detected characters plus false positives as represented in Equation (4.10).

Definition 2: False positives (FP) / False alarms are those regions in the image which are actually not characters of a text, but have been detected by the algorithm as a text.

\[
\text{Precision rate} = \frac{\text{Correctly detected characters (True positives)}}{\text{Correctly detected characters + False positives}} \times 100\% \tag{4.10}
\]

Definition 3: The Recall rate (R) is defined as the ratio of the correctly detected characters to the sum of correctly detected characters plus false negatives as represented in Equation (4.11).

Definition 4: False negatives (FN)/ Misses are those regions in the image which are actually text characters, but have not been detected by the algorithm.

\[
\text{Recall rate} = \frac{\text{Correctly detected characters}}{\text{Correctly detected words + False negatives}} \times 100\% \tag{4.11}
\]
**Definition 5:** The F-score is the harmonic mean of the recall and precision rates as represented in Equation (4.12).

\[
F\text{-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{Precision} + \text{recall}}
\] (4.12)

### 4.4.3 Experimental Results

The system has been tested on the oriented text in horizontal and vertical directions with mixed languages (English and Tamil) and one such example is shown in Figure 4.8. Text extraction from a Web document image is shown in Figure 4.9. The output image of the proposed algorithm in Figures 4.9 to 4.11 consists only of the detected text for the document image, scene text and caption text image respectively.

Figure 4.8 Results of oriented text with mixed languages

Figure 4.9 Results of Web document image
The results of the experiments of three types of sample images are presented in Table 4.2, where the number of True positives, number of false alarms/false positives, number of False negatives/Misses and the corresponding values for precision, recall and F-score are listed. The average of all measures for the three types of images have been listed in Table 4.3 and it is observed from Table 4.3 that

- The proposed algorithm produces the highest recall rate for the caption text image as only some relatively weak texts are missed, and shows a promising precision rate for document images, as non-text objects are correctly classified as shown in Figure 4.12.

- The recall rate is comparable for the three types of images.
• It is found from experiments that the Precision rate, and in turn the average F-score of the scene text images dropped when compared to the caption text images and document images, because of the embedded nature of the text inside the image as shown in Figure 4.13. Experimentation is required to bring out the performance of the extraction of the scene text image comparable to the caption text and document images.

Table 4.2 Performance measures for the SBTA-TD / TL method (for a dataset with sample images)

<table>
<thead>
<tr>
<th>Img type</th>
<th>Img no</th>
<th>TP/ Detections</th>
<th>FP/ False alarms</th>
<th>FN/ Misses</th>
<th>P</th>
<th>R</th>
<th>F-SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scene text images</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scn 1</td>
<td>30</td>
<td>3</td>
<td>2</td>
<td>90.9</td>
<td>93.7</td>
<td>92.3</td>
<td></td>
</tr>
<tr>
<td>Scn 2</td>
<td>18</td>
<td>6</td>
<td>2</td>
<td>75</td>
<td>90</td>
<td>81.8</td>
<td></td>
</tr>
<tr>
<td>Scn 3</td>
<td>6</td>
<td>5</td>
<td>1</td>
<td>54.5</td>
<td>85.7</td>
<td>66.6</td>
<td></td>
</tr>
<tr>
<td>Scn 4</td>
<td>50</td>
<td>12</td>
<td>-</td>
<td>80.6</td>
<td>100</td>
<td>89.3</td>
<td></td>
</tr>
<tr>
<td>Scn 5</td>
<td>24</td>
<td>19</td>
<td>-</td>
<td>38.7</td>
<td>100</td>
<td>55.8</td>
<td></td>
</tr>
<tr>
<td>Caption text images</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cap 1</td>
<td>49</td>
<td>2</td>
<td>-</td>
<td>94.2</td>
<td>100</td>
<td>97</td>
<td></td>
</tr>
<tr>
<td>Cap 2</td>
<td>23</td>
<td>8</td>
<td>-</td>
<td>74.2</td>
<td>100</td>
<td>85.2</td>
<td></td>
</tr>
<tr>
<td>Cap 3</td>
<td>51</td>
<td>6</td>
<td>-</td>
<td>89.4</td>
<td>100</td>
<td>94.4</td>
<td></td>
</tr>
<tr>
<td>Cap 4</td>
<td>45</td>
<td>5</td>
<td>-</td>
<td>90</td>
<td>100</td>
<td>94.7</td>
<td></td>
</tr>
<tr>
<td>Cap 5</td>
<td>36</td>
<td>6</td>
<td>-</td>
<td>85.7</td>
<td>100</td>
<td>92.3</td>
<td></td>
</tr>
<tr>
<td>Document Images</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Docpg 1</td>
<td>134</td>
<td>4</td>
<td>5</td>
<td>97</td>
<td>96</td>
<td>96.5</td>
<td></td>
</tr>
<tr>
<td>Docpg 2</td>
<td>45</td>
<td>4</td>
<td>4</td>
<td>92</td>
<td>92</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td>Docpg 3</td>
<td>74</td>
<td>2</td>
<td>52</td>
<td>97</td>
<td>58.7</td>
<td>73.1</td>
<td></td>
</tr>
<tr>
<td>Docpg 4</td>
<td>39</td>
<td>2</td>
<td>29</td>
<td>95</td>
<td>57.3</td>
<td>71.5</td>
<td></td>
</tr>
<tr>
<td>Docpg 5</td>
<td>684</td>
<td>3</td>
<td>10</td>
<td>99.5</td>
<td>98.5</td>
<td>99</td>
<td></td>
</tr>
</tbody>
</table>
Table 4.3 Average Performance measures for the SBTA –TD/TL system

<table>
<thead>
<tr>
<th>Measures</th>
<th>Image sets</th>
<th>Set 1 (Caption text images)</th>
<th>Set 2 (Scene text images)</th>
<th>Set 3 (Document images)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of True positives (TP)</td>
<td>Set 1</td>
<td>37</td>
<td>31.36</td>
<td>257.66</td>
</tr>
<tr>
<td>Number of False positives (FP)</td>
<td>Set 2</td>
<td>8.9</td>
<td>7.45</td>
<td>3.66</td>
</tr>
<tr>
<td>Number of False negatives (FN)</td>
<td>Set 3</td>
<td>5.71</td>
<td>13.66</td>
<td>18.55</td>
</tr>
<tr>
<td>Recall</td>
<td>Set 1</td>
<td>91.7 %</td>
<td>89.74 %</td>
<td>87 %</td>
</tr>
<tr>
<td>Precision</td>
<td>Set 2</td>
<td>81.5 %</td>
<td>72.5 %</td>
<td>97 %</td>
</tr>
<tr>
<td>F-Score</td>
<td>Set 3</td>
<td>86 %</td>
<td>78 %</td>
<td>91 %</td>
</tr>
</tbody>
</table>

Figure 4.12 Precision and Recall bar chart of SBTA –TD/TL method

Figure 4.13 F-Score line graph of SBTA –TD/TL method
4.4.4 Comparison with other Text Extraction Techniques

To give an average estimate of the performance of the TD / TL system, the results have also been verified against two other existing algorithms: ie) the edge based method (Xiaoqing and Jagath 2005, 2006) and the connected component (CC) method (Julinda et al 2003). The results show that the proposed system shows a clear improvement over the edge and CC methods as shown in Figures 4.14 and 4.15 for scene text and caption text respectively.

(a) (b)

(c) (d)

Figure 4.14 Comparison of SBTA –TD /TL and various methods for
Scene text image
a) Scene text. b) Edge based method c) CC method d) Proposed
SBTA method
Even though the edge based method (Xiaoqing and Jagath 2005, 2006) is designed for scene text and document images and the CC method (Julinda et al 2003) is designed for caption text images, the proposed method is compared with the above two methods in order to show the ability of the proposed method to extract a text from all the three kinds of images in a better way, as there is no single methodology available to compare. The Precision, Recall and F-Score are estimated using the proposed method, Edge based and CC methods with the subset of our data sets. The performance measures for comparison are shown in Tables 4.4 and 4.5. It is observed from these tables that

i) **Caption text images (CTI):** The proposed method produces 81.5 % of precision, 91.7% of recall and 86% of F-score which is an encouraging result when compared to the CC method which is designed for the caption text image. The CC
method extracts a text with an F-score of 75% and also gives poor character quality for all images; this is shown in Figures 4.14(c) and 4.15(c). The edge based method is able to produce an F-score of 66% only.

ii) **Scene text images (STI):** The proposed method produces an F-score of 78% which is better than the 63% of the edge based method, which was designed to handle scene text images. The CC method produces a very low performance of 34% F-score with poor produced character quality.

iii) **Document images (DI):** The proposed method produces 97% of precision and 87% of recall with an F-score of 91%, which is a promising result when compared to both the CC and Edge methods.

So, it is concluded from the results that, in all cases, the proposed method is found to be better than the CC method and the edge based method for all three kinds of images by providing better precision, recall and F-score rates. This is shown in Figures 4.16 and 4.17 as a comparison of the 3-D cube and line graph.

**Table 4.4 Comparison of Average Performance measures for subset of dataset for SBTA-TD/TL and other methods**

<table>
<thead>
<tr>
<th>Img type</th>
<th>Measures</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F-Score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Edge</td>
<td>CC</td>
<td>Proposed method</td>
</tr>
<tr>
<td>Caption text images</td>
<td></td>
<td>62</td>
<td>69</td>
<td>81.5</td>
</tr>
<tr>
<td>Scene text images</td>
<td></td>
<td>61</td>
<td>44</td>
<td>72.5</td>
</tr>
<tr>
<td>Document images</td>
<td></td>
<td>74.8</td>
<td>80.8</td>
<td>97</td>
</tr>
</tbody>
</table>
Table 4.5 Average F-score for the SBTA-TD/TL and other methods

<table>
<thead>
<tr>
<th>Image type</th>
<th>Edge based Method</th>
<th>CC method</th>
<th>Proposed method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caption text images</td>
<td>66 %</td>
<td>75 %</td>
<td>91%</td>
</tr>
<tr>
<td>Scene text images</td>
<td>63 %</td>
<td>34 %</td>
<td>78%</td>
</tr>
<tr>
<td>Document images</td>
<td>71.6 %</td>
<td>75 %</td>
<td>91%</td>
</tr>
</tbody>
</table>

Figure 4.16 3D cube for F-score comparison of the SBTA-TD/TL and other methods

Figure 4.17 Line graph for F-score comparison of the SBTA-TD/TL and other methods
4.5 SUMMARY

In this section, a novel SBTA based text detection/localization (SBTA –TD/ TL) method is presented. The gradients of the contours are used to detect the text by combining the strong and boosted weak sub bands. The algorithm has been applied on several images of three types with complex backgrounds; encouraging results were obtained and the experimental results show that the proposed method outperforms the edge based method and the connected component method. The results indicate that our methodology using the Non sub sampled contourlet transform as a texture based method has the efficacy to discriminate between a text and a non text for three kinds of images. During the experiments, it was observed that

- The SBTA based TD/TL method works across different kinds of images such as Caption text images, Scene text images and Document images.

- This is robust to limited range of font size of characters.

- This is robust to orientation of text string (Horizontal, Vertical and diagonal )

- This method can be easily extended to other languages.

It was also observed that the SBTA – TD/TL method does not show encouraging results for the following variations

- Variation in lighting conditions (Edge contrast changes according to the brightness of the image which is taken in day light and night light.)

- Variation in a wide range of Font sizes (Font size above 30 and below 12)
- Variation in the orientation of the text (Inclined text or text on a circular arc or radially changing text)

- Perspective projection (images in 3-D space, where characters can be distorted by slant, tilt and shape of objects on which the characters are printed).

Consequently, all these observations, drawbacks and advantages were analyzed, and in the next chapter a more complex method focusing on the above challenges will be proposed.