Abstract—In this paper, we address curved text detection in video through a new enhancement criterion and the use of quad tree. The proposed method makes use of the quad tree to simplify the task of handling the entire frame at each stage. The proposed method employs a novel criterion for grouping of pixels based on their R, G and B values to enhance text information. As generally, a text detection problem is a two class problem, we used k-means with k=2 to identify potential text candidate pixels. From these potential candidates, connected components are then extracted and subjected to further analysis, where symmetry property based on stroke width is used for further authentication of the text representatives. These authenticated text representatives are then exploited as seed points to restore the text information with reference to the Sobel edge frame of the original input frame. To preserve the spatial information of text pixels the concept of quad tree is applied. From these seed blocks, text lines are extracted by the use of a region growing approach driven completely based on Sobel edge map. The proposed method is tested on curved video data and Hua’s horizontal video text data in terms of recall, precision, f-measure, misdetection rate and processing time. The results are compared and analyzed to show that the proposed method outperforms several existing methods in terms of accuracy and efficiency.

Keywords—Curved text detection; Text enhancement; k-means clustering; Quad tree technique; Symmetry verification

I. INTRODUCTION

Curved text detection and extraction from a video frame is more challenging than horizontal and non-horizontal text lines detection due to non-linear orientation of characters and words in curved text lines compared to linear orientation of characters and words in a horizontal and non-horizontal text lines. Nowadays video capturing of day to day activities is becoming more common with advancement of video technology. As a result, video databases increase drastically in size. Therefore, developing efficient and accurate methods through text detection and recognition for retrieving desired data from a huge database is receiving great interests [1-5]. Thus there is a demand for developing methods which can detect text without any assumptions and restrictions.

There are plenty of methods for horizontal [6-10] and non-horizontal [11-15] text detection in the literature but hardly one method can be found [16] for curved text detection in the literature because of its complexity. For example, Zhou et al. [11] have proposed a method based on connected component analysis to detect horizontal and vertical text lines using color information. The scope of this work is limited to horizontal and vertical text lines and thus not applicable to curved text lines.

Yao et al. [12] have proposed a method for arbitrary text detection in scene images by extending stroke width transform [13] and with the help of classifiers. However, this method considers non-horizontal text lines as arbitrary text lines. But in theory, “arbitrary” should mean inclusion of curved text lines. Recently, Shivakumara et al. [14, 15] have developed a number of methods by exploiting Laplacian and Fourier transform combination with skeleton concept, and Bayesian classifier with boundary growing to detect non-horizontal text lines. These methods do not consider curved text lines for experimentation. However, the same authors [16] have developed another method based on gradient directions and two stage boundary growing to address the problem of arbitrary text lines including curved text. The main drawback of this method is the use of a classification algorithm which classifies given input data as horizontal and non-horizontal text lines (including curved text lines) using angle information of words to apply boundary growing. In addition, the classification algorithm depends on word segmentation to classify the frame as having horizontal text lines or non-horizontal text lines. It is not clear how the algorithm will do with a frame that contains both horizontal and non-horizontal text lines. Overall, the method is computationally expensive because of the two stage connected component analysis and growing process. These factors motivated us to propose a new method to overcome the above problems by exploring new grouping, symmetry criteria, quad tree and boundary growing in this work.

II. PROPOSED METHODOLOGY

The proposed method is structured into four sub-sections. In section A, a grouping criterion based on the observation that text pixels have high intensity values than non-text pixels in R, G and B bands of the input video frame due to high contrast of text pixels compared to its background. In addition, grouping according to neighbor values of pixels using color values is also presented to enhance the text information. Since grouping helps in widening the gap between text and non-text pixels, we propose a k-means with k=2 clustering algorithm to classify text pixels which we call potential text candidates. Symmetry criteria based on stroke width transform for skeleton of the potential text candidates is presented in section B to verify further potential text candidates. Each authenticated potential text candidate is used to restore the text information with the help of Sobel edge map of the input frame, which we call text representatives. In order to overcome problems due to non-text representatives, we introduce a quad tree technique to preserve the spatial information of text pixels to identify the seed text
blocks in Section C. Subsequently, boundary growing is proposed using seed blocks and by referring Sobel edge image to extract curved text in the frame in Section D.

A. Grouping Color Values for Text Enhancement

As we are inspired by the work presented in [10] where it is shown that color information in R, G and B is useful for enhancing text information, we propose a new grouping criterion to enhance the text information by suppressing non-text information in this work. The basis for this grouping is that text pixels have high color values compared to background values. For the color input frame as shown in Fig. 1(a) where the image contains curved text with complex background, the method obtains three color sub-bands frames, namely R, G and B. In order to identify the high color value in three sub-bands, we use a simple Max-Min clustering criterion which selects Maximum (Max) and Minimum (Min) value from R, G, B sub-bands for each pixel. Then the third value is compared with Max and Min values to find its closest value. If the third value is closer to Max value then it forms a Max cluster with Max value else it forms a Min cluster with Min value. If the third value forms a Max cluster then the method chooses Max value in the Max cluster to replace actual pixel value else it chooses Min value in the Min cluster to replace actual pixel value. In this way, the Max-Min clustering helps in grouping to identify the high color value for each pixel in the input image which results in an enhanced image as shown in Fig. 1(b) where one can notice text pixels are enhanced and non-text pixels are suppressed.

The previous step shows that text pixels have high color values among R, G and B sub-bands of the input frame. This clue motivates us to check neighboring text pixels’ values in the enhanced image to choose high value among its neighbors by applying the same grouping criteria to enhance text further. We propose a sliding window (3×3) operation where we use the above process to sharpen the text pixels and to suppress the non-text pixels based on neighbor information further as shown in Fig. 1(c) where the text pixel are still brighter than the pixel in the enhanced image shown in Fig. 1(b). The size of the sliding window is determined based on experimental study.

(b) Enhanced image
(c) Sharpened image

Fig. 1. Potential candidate text pixels selection through grouping criteria

B. Symmetry Criteria for Text Representative

It is noticed from Section II.A that the enhancement steps increases the gap between text and non-text pixels. In order to separate text pixel, we propose to employ k-means clustering algorithm with k=2 since it is a two class problem. We consider the cluster that gives a high mean value compared to the other cluster as a text cluster. Though the k-means clustering algorithm classifies the text cluster, it may still contain non-text pixels due text like features in the background, which results in text candidates. Therefore, we propose a connected component analysis based on symmetry criteria to filter out non-text components from text candidates. The method reduces pixel width of text candidates to single pixel using a skeleton concept as shown in Fig. 1(d) where due to low resolution and complex background, skeleton may not preserve the shape of the components but it gives significant information to study the characteristics of text and non-text components. To remove non-text components from the skeleton image, we test whether the text components satisfy fully the connected component condition or not because it is a fact that text components must be connected without any disjoint compared to non-text components where most of them are disjoint. The output of performing a fully connected component testing is shown in Fig. 1(e) where one can see most of the non-text components are removed and we call this output potential text candidates image. The fully connected component is defined as one whose starting and ending of its contour should meet at one point. However, we still can see some non-text components in the results shown in Fig. 1(e).

The symmetry criterion is proposed based on the fact that the stroke width is constant throughout the character while for non-text the stroke width distance is arbitrary. For each pixel in the fully connected component, the method computes the stroke width distance [13] that is traversing in the perpendicular direction to the gradient direction of the pixel till it reaches a white pixel which we call a reached pixel. In this way, the stroke width is computed for each pixel in the fully connected component. Then to find the dominant stroke width distance, we plot a histogram for the stroke width distances to choose the stroke width distance which gives the highest peak as the dominant stroke width distance. Due to low resolution and complex background, it is hard to get a complete shape of the character. Therefore, one cannot expect a constant stroke width for the character and hence we choose the dominant stroke width distance for verification. For each reached pixel of the component, we obtain the dominant stroke width distance. Then the method compares those two dominant stroke width distances to test the symmetry criteria. If both the distances are the same then it is considered a text component as it satisfies the symmetry criterion. This is true because according to the stroke width concept presented in [13], the stroke width distance of the starting pixel and the reached pixel should be same. Note that the gradient image is obtained by performing the vertical and horizontal mask operation on the enhanced image shown in Fig. 1(b). The result of symmetry verification is shown in Fig. 2(a) where almost all non-text components are removed. Fig. 2(a) shows that the symmetry verification removes sometimes text components. The method extracts edge components in the Sobel edge image of the input frame shown in Fig. 2(b) corresponding to the components in the output of symmetry verification to restore the missing text.
information as shown in Fig. 2(c) where it is noticed that text components are restored and non-text components are removed. This output is called an image with text representatives.

C. Quad Tree Technique for Seed Text Blocks

Based on experimental results of the methods presented in Section II.B, we observe that preventing non-text components completely is not possible as shown in Fig. 2(c) where text like object is still exists in the image because of complex background. It is also true that text detection methods usually propose heuristics based on geometrical properties of components to eliminate text like objects (non-text components). However, studying characteristics at the component level is not advisable as it requires more computational time. These factors motivated us to propose a quad-tree technique to identify the seed blocks from the text representative image because quad-tree works at the block level but not at the component level. Therefore, it reduces computational burden. The method checks if a divided block contains text components or not. If the divided block does not contain text then the method does not apply the quad tree technique. In this way, we use the quad-tree technique efficiently in this work. We divide the whole text representative image into four equal parts. The method explores the spatial relationship between the components in each block to extract regular spacing between the text components. The method finds the centroid of each component first. Next we find distances among these centroids as follows. From the first centroid c1 find out the nearest centroid c2 then find the distance between them. Now find the nearest centroid for c2, say c3 and the distance between them. In this way, we found the distance between centroids of all the components in the division. We apply k-means (k=2) on distances to choose the cluster that gives the lower mean between the two. Then we will check the standard deviation of distances before clustering and after clustering blocks. If the standard deviations of both are not equal then again we apply quad tree technique to divide the corresponding block into four equal parts otherwise we stop dividing the blocks. The basis for this is that if there is a text then the standard deviation of distances before clustering and after clustering will be almost the same or else there will be much variation. The output of the quad-tree technique is considered as seed blocks as shown in Fig. 3 where seed blocks for each iteration and the final combined all seed blocks as one block are shown. Note that the iterative process continues until it becomes of size 32x32. More details for conditions to select seed blocks are as follows.

\[ Diffstd = abs(BST - AST) \] (1)

\[ NC = BC/AC \] (2)

\[ NP = BP/AP \] (3)

\begin{align*}
\text{In equation (1), } & \text{Diffstd is the absolute value of the difference of before and after cluster standard deviation. BST is the standard deviation of before cluster and AST is the standard deviation of after cluster block. In equation (2) NC is the number of components. BC is the number of components present before clustering and AC is the number of components present after clustering. NP is the number of pixels. BP is the number of pixels present before clustering and AP is the number of pixels present after clustering. Based on these three equations we derive a dynamic threshold equation. }
\end{align*}

\[ Diffstd >= NC/NP \] (4)

If equation (4) is satisfied then it is considered as a seed block otherwise it is a non-text block. If the standard deviation before and after clustering the block is that same then directly we considered that as a seed block otherwise we apply the quad tree method and we check that the above four conditions to select the seed blocks. If this result still contains some false positives then we propose simple rules to eliminate them.

D. Region Growing for Curved Text Extraction

The seed block given by the quad-tree technique is considered for full text line extraction by proposing a region growing method which grows components in the seed block pixel by pixel until it reaches a neighboring component along the text direction in the Sobel edge map of the enhanced image. This process continues till the end of the text line. The end of the text line is determined by comparing the spaces between text lines, words and characters dynamically. The region growing works based on the fact that the space between text lines is greater than the space between words and between characters. The main advantage of the region growing is that it extracts text line of any orientation (such as curved texts) because it works based on the nearest neighbor concept. For
example, the nearest neighbor for the first character in the text line would be the second character and for the second character, the third character is the nearest neighbor. The final output of region growing is shown in Fig. 4 where the text line is extracted with a false positive for the image shown in Fig 1(a).

Fig. 4. Text extraction through boundary growing

III. EXPERIMENTAL RESULTS

We conduct experiments on arbitrary data used in [6] for curved text extraction and one more standard dataset used in [17] for horizontal text detection. The arbitrary data consists of 142 curved text images and Hua’s data [17] consists of 45 horizontal text images. To evaluate the proposed method, we use Recall, Precision (P), F-measure (F) and Misdetection Rate (MDR) and Average Processing Time (APT) in seconds. In summary, 187 video frames (142+45) are used for purpose of evaluation. To show that the proposed method is superior to existing methods, we compare recent methods of horizontal, non-horizontal and arbitrary texts. Horizontal text detection methods such as Cai et al’s [19] method uses color and connected component analysis, Wong and Chen’s [7] method uses gradient and statistical information of seed text lines, Liu et al’s [6] method proposed texture features with k-means clustering, Fourier-RGB method [10] combines Fourier and color information. Non horizontal text detection methods such as Zhou et al.’s [11] method use color and effective component analysis for caption and big font text detection, Laplacian method [14] works well for non-horizontal by combining Laplacian and Fourier, Bayesian method [15] explores Bayesian classifier and gradient. We found only one method for arbitrary text detection by Sharma et al. [16] which proposes gradient directions and two stage grouping. However, this method is not robust to handle curved text and horizontal text when the image contains both of them since the classification method fails to classify such images correctly. In addition, the classification method requires word segmentation.

We define the following measures for each detected text block to evaluate the performance of the proposed method, as it is widely used in literature [14, 15, 11, 10, 6, 7].

Truly Detected Block (TDB): A detected block that contains at least one true character. Thus, a TDB may or may not fully enclose a text line. Falsely Detected Block (FDB): A detected block that does not contain text. Text Block with Missing Data (MDB): A detected block that misses more than 20% of the characters of a text line (MDB is a subset of TDB). The percentage is chosen according to [14], in which a text block is considered correctly detected if it overlaps at least 80% of the ground-truth block. We count manually Actual Number of Text Blocks (ATB) in the images and it is considered as ground truth for evaluation.

The performance measures are defined as follows. Recall (R) = TDB / ATB, Precision (P) = TDB / (TDB + FDB), F-measure (F) = (2PR) / (P + R), Misdetection Rate (MDR) = MDB / TDB.

A. Experimental Results on CurvedData

Sample results of the proposed method for curved text extraction are shown in Fig. 5 where (a) shows input frames having different background and fonts and (b) shows the results given by the proposed method for corresponding images shown in (a). Fig. 5 shows that the proposed method extracts curved text successfully with few false positives. We need to investigate how to eliminate false positives completely. The quantitative results of the proposed and existing methods are reported in Table I. Table I show that the proposed method outperforms the existing methods in terms of recall, precision and F-measure while Fourier-RGB method is the best in MDR and worst in recall compared to the proposed method. Wong and Chen method is the best in APT but worst in recall compared to the proposed method. The main reason for poor results of the existing methods is that the existing methods (Bayesian, Laplacian, Zhou et al, Fourier-RGB, Liu et al, Wong and Chen, Cai et al) are developed for horizontal and non-horizontal text detection but not for curved text line detection. However, Sharma et al method is developed for arbitrary text detection but it has its own inherent problems to achieve better accuracy. On the other hand, the proposed method is developed for any kind of text detection without assumptions and constraints as in the existing methods. Therefore, the proposed method gives better results compared to existing methods.

<table>
<thead>
<tr>
<th>Methods</th>
<th>R</th>
<th>P</th>
<th>F</th>
<th>MDR</th>
<th>APT (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed Method</td>
<td>0.80</td>
<td>0.83</td>
<td>0.81</td>
<td>0.25</td>
<td>2.3</td>
</tr>
<tr>
<td>Sharma et al[16]</td>
<td>0.73</td>
<td>0.88</td>
<td>0.79</td>
<td>0.28</td>
<td>10.3</td>
</tr>
<tr>
<td>Bayesian [15]</td>
<td>0.59</td>
<td>0.52</td>
<td>0.55</td>
<td>0.27</td>
<td>12.1</td>
</tr>
<tr>
<td>Laplacian [14]</td>
<td>0.55</td>
<td>0.68</td>
<td>0.60</td>
<td>0.42</td>
<td>9.9</td>
</tr>
<tr>
<td>Zhou et al. [11]</td>
<td>0.41</td>
<td>0.60</td>
<td>0.48</td>
<td>0.38</td>
<td>2.2</td>
</tr>
<tr>
<td>Fourier-RGB [10]</td>
<td>0.52</td>
<td>0.68</td>
<td>0.58</td>
<td>0.16</td>
<td>17.5</td>
</tr>
</tbody>
</table>

Fig. 5. Sample curved text extraction results of our method
B. Experimental Results on Horizontal Data (Hua’s Data)

Sample results of the proposed method for horizontal text detection are shown in Fig. 6 where (a) shows few frames of different kinds of text and (b) shows the results of the proposed method for the corresponding images in (a). Fig. 6(b) shows that the proposed method extracts horizontal text lines properly with some false positives. Hua’s data is considered as a benchmark data for experimentation in this work as it is the only data available publicly at (http://www.cs.cityu.edu.hk/~liuwy/PE_VTDetect/).

The quantitative results of the proposed and existing method are reported in Table II. Table II shows that the proposed method is comparable but not the best in terms of recall, precision and F-measure. However, the proposed method gives low MDR and processing time compared to all the existing methods. This is because of the advantage of quad-tree technique on text representatives. The proposed method is better than Sharma et al. in terms of F-measure, MDR and APT. The Laplacian and Bayesian methods give better recall, precision and F-measure than the proposed method because these methods take full advantage of horizontal text information while the proposed method works for curved text images without sacrificing much for horizontal text detection.

(b) Horizontal text lines extraction corresponding to the above input frames shown in (a)

TABLE II. PERFORMANCE OF THE PROPOSED AND EXISTING METHODS ON HORIZONTAL (HUA’S) DATA

<table>
<thead>
<tr>
<th>Methods</th>
<th>R</th>
<th>P</th>
<th>F</th>
<th>MDR</th>
<th>APT (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed Method</td>
<td>0.82</td>
<td>0.88</td>
<td>0.84</td>
<td>0.06</td>
<td>2.0</td>
</tr>
<tr>
<td>Sharma et al. [16]</td>
<td>0.88</td>
<td>0.77</td>
<td>0.82</td>
<td>0.32</td>
<td>9.0</td>
</tr>
<tr>
<td>Bayesian [15]</td>
<td>0.87</td>
<td>0.85</td>
<td>0.85</td>
<td>0.18</td>
<td>5.6</td>
</tr>
<tr>
<td>Laplacian [14]</td>
<td>0.93</td>
<td>0.81</td>
<td>0.87</td>
<td>0.07</td>
<td>11.7</td>
</tr>
<tr>
<td>Zhou et al. [11]</td>
<td>0.72</td>
<td>0.82</td>
<td>0.77</td>
<td>0.44</td>
<td>1.13</td>
</tr>
<tr>
<td>Fourier-RGB [10]</td>
<td>0.81</td>
<td>0.75</td>
<td>0.76</td>
<td>0.06</td>
<td>14.6</td>
</tr>
<tr>
<td>Liu et al. [6]</td>
<td>0.75</td>
<td>0.54</td>
<td>0.63</td>
<td>0.16</td>
<td>24.9</td>
</tr>
<tr>
<td>Wong and Chen [7]</td>
<td>0.51</td>
<td>0.75</td>
<td>0.61</td>
<td>0.13</td>
<td>1.6</td>
</tr>
<tr>
<td>Cai et al. [9]</td>
<td>0.69</td>
<td>0.43</td>
<td>0.53</td>
<td>0.13</td>
<td>9.2</td>
</tr>
</tbody>
</table>

IV. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a new grouping criterion based on color information in sub-bands and neighboring text pixels’ information in the enhanced image to sharpen text pixels and to widen the gap between text and non-text pixels. A new symmetry criterion is introduced based on stroke width transform to filter out non-text components. Seed blocks from text representatives are obtained by the quad-tree technique which uses spatial information of text pixels. To the best of our knowledge, quad-tree technique for text detection is the first attempt. Finally, region growing is proposed to extract full text lines with the help of Sobel edge image of the input frame. Future work will aim to eliminate false positives without using heuristics and exploring learning based methods.

ACKNOWLEDGEMENT

This work is done jointly by University of Malaya, National University of Singapore and University of Mysore. It is also supported in part by A*STAR grant 092 101 0051 (WBS no. R252-000-402-305).

REFERENCES