Multi-Spectral Fusion based Approach for Arbitrarily-Oriented Scene Text Detection in Video Images

Guozhu Liang, Palaihnakote Shivakumara, Tong Lu, Member IEEE and Chew Lim Tan, Senior Member, IEEE

Abstract—Scene text detection from video as well as natural scene images is challenging due to the variations in background, contrast, text type, font type, font size, etc. Besides, arbitrary orientations of texts with multi-scripts add more complexity to the problem. The proposed approach introduces a new idea of convolving Laplacian with wavelet sub-bands at different levels in the frequency domain for enhancing low resolution text pixels. Then the results obtained from different sub-bands (spectral) are fused for detecting candidate text pixels. We explore maxima stable extreme regions along with stroke width transform for detecting candidate text regions. Text alignment is done based on the distance between the nearest neighbor clusters of candidate text regions. Additionally, the approach presents a new symmetry driven nearest neighbor for restoring full text lines. We conduct experiments on our collected video data as well as several benchmark datasets such as ICDAR 2011, ICDAR 2013 and MSRA-TD500 to evaluate the proposed method. The proposed approach is compared with the state of the art methods to show its superiority to the existing methods.

Index Terms—Laplacian-wavelet, Multi spectral fusion, Maxima stable extreme regions, Stroke width transform, Arbitrarily oriented video text detection.

1. INTRODUCTION

With the recent advances in multimedia and network technologies combined with the rapid decline in hardware prices, the contents of digital video and images are growing at a tremendous speed. As the statistics data of 2010 shows, more than 35 hours of video contents were uploaded to video sharing sites (e.g., YouTube) every minute and more than 35 million photos had been uploaded to social networking sites like Facebook every month [1]. This results in huge databases and thus requires approaches which work at high level semantics. The conventional approaches that use low level features may not be sufficient for handling such large databases due to the gaps between low level features and high level semantics. To alleviate this problem, text detection and recognition has become popular as it provides meaningful cues which are close to the content of video or image [2-4]. So, it has been widely used in video summarization, content based image indexing and video sequence retrieval. On top of these applications, text detection and recognition has also been used for real time surveillance applications, such as assisting a blind person to walk freely on roads, assisting tourists to reach their destinations, enhancing safe driving, navigating vehicles based on license plate information, exciting event extraction from sports video, identifying athletes in marathon events, etc [5].

Video consists of two types of texts, namely, caption text and scene text. Caption text is manually edited, which has good clarity and visibility and hence is easy to process. Scene text exists naturally in video frames, the detection of which suffers from color bleeding, low contrasts, low quality due to distortion, different orientations, backgrounds, etc. Hence, scene text is hard to process compared to caption text [4, 6, 7].

Scene images captured through a high resolution camera usually contain only scene texts with high contrast and complex background, while video contains both caption and scene texts with low resolution and complex background. Achieving a good accuracy for text detection from both video and natural scene images is still an open issue for researchers in the field of image processing and pattern recognition because most of the existing approaches [8, 9] either focus on caption text in video or scene text in natural scene images but not both video and natural images.

The problem of text detection and recognition from scanned document images is not new for the document analysis community because for different scripts we can find several Optical Character Recognizers (OCR engines) that are available publicly. However, the same methods may not be used for detection and recognition of the texts in video and natural scene images because the approaches work well for plane background and high contrast images but not for images like video and natural scene images [10-12].

To widen the scope of document analysis based approaches, there are methods proposed for text detection from natural scene images [13-18]. These approaches directly or indirectly rely on the features of connected components and the shapes of characters to achieve a good accuracy. This is valid because text in natural scene images usually has high contrast as mentioned above and hence the shape of a character can be preserved in most of the situations. However, this is not
necessarily true for video, where we can expect disconnections, loss of information, distorted shapes and so on due to low resolution and low contrast. Therefore, the approaches developed for text detection in natural scene images may not be used directly for text detection in video.

There are approaches [19-25] proposed for text detection by exploring temporal frames. Generally, these approaches use temporal frames to produce fine details of texts and false positive elimination by integrating temporal frames into single enhanced frame. The big question lies in finding the exact number of frames for the enhancement process. Currently, the approaches use a fixed number of frames as it is hard to validate the output of the enhancement process. In addition, these approaches work only for video rather than an individual image.

In summary, developing a method that works well for text detection in both video and natural scene images is challenging, and hence it is a research issue worth exploring.

II. RELATED WORK

For text detection in video, a large number of approaches have been developed. The approaches can be classified broadly into three categories: connected component based [11, 12], edge and gradient based [8, 9, 26], and texture based approaches [6, 7, 20].

Connected component based approaches are fast and good for images that have high contrast texts and plain background just like methods in the document analysis field. On the other hand, these approaches will not be suitable for text detection in video and natural scene images due to low video resolution and natural scene complexity. To improve the performance of text detection, edge and gradient based approaches are developed. These approaches are good at recall but poor at precision because the proposed features are sensitive to background complexity leading to more false positives.

Texture based approaches are developed to solve the problems of edge and gradient based approaches because the texture property works well for complex background. Here, the approaches define the appearance of text pattern as a special texture. For example, Shivakumara et al. [6] proposed a Laplacian approach for multi-oriented text detection in video based on the combination of Laplacian and Fourier transform. This approach uses this combination for enhancing low contrast texts in video. With the help of k-means clustering, the approach obtains text candidates. Skeleton concept has been used for segmenting text lines based on end and junction points. However, this approach is developed for text detection in video and gives low precision due to the confusion to determine exact end and junction points especially for Chinese script. Li et al. [20] proposed an approach for text detection and tracking in video using the combination of wavelet and moments. The combination is good for horizontal text detection. However, the approach is not tested on natural scene images to show whether it is effective for scene images.

Liu et al. [27] proposed a novel approach for multi-oriented Chinese text detection in video using wavelets. The color and spatial relationship of the structure of text candidates are used for text candidates merging. The SVM classifier is proposed for false positive elimination. The method uses language specific features for text detection. Zhang et al. [28] proposed video text extraction using fusion of color, gradient and log-Gabor filter. The approach fuses color, gradient and log-Gabor images to enhance text information in video frames. Then the approach proposes character segmentation based on vertical profiles before text extraction. However, this approach is limited to horizontal text but not arbitrarily oriented text of different scripts.

Li et al. [16] introduces the so-called characterness for text detection in natural scene images based on finding saliency to measure the so-called objectness. The approach explores a Markov field random model to exploit the inherent dependencies between characters. Finally, a Bayesian approach is proposed for the integration of the characters obtained from different clues. The approach misclassifies characters sometime because of the characterness that exists in the background. In addition, the main focus of this approach is to detect texts from natural scene images but not video. Yi et al. [9] proposed text localization in scene images using clustering, stroke segmentation and string fragment classification. The approach combines structural analysis of text stroke with color assignment to filter out background interferences. A robust string fragment classification method is proposed using Gabor-based text features. However, the approach requires high resolution to achieve a good accuracy.

Zhang and Kasturi [8] proposed a novel text detection system based on character and link energies. The approach first extracts connected components from an edge image and then proposes to use characteristics of text characters to link neighboring characters. Finally, it combines the characters and links energies to compute text unit energy as a measure of the likelihood of the candidate being a text object. However, the performance of the approach depends on the edge image of the input image.

Similarly, Shivakumara et al. [26] proposed the use of gradient vector flow and a grouping based approach for arbitrarily oriented text detection in video frames. The approach explores the gradient vector flow at the corners of text components in an edge image for identifying text candidates. The two stage grouping is proposed to extract arbitrarily oriented texts in video. This approach is good for non-horizontal but not for horizontal text lines because the proposed grouping merges two text lines as one text line due to the variations in the spacing between two text lines.

In light of the above discussions, we can notice that connected component based approaches are good for studying geometrical features of text components, edge and gradient based approaches are good for finding inter and intra characters symmetry, while texture based approaches are good for text detection from complex backgrounds. Besides, we believe that the texture patterns of different texts such as caption and scene text from video and natural scene images appear similarly. These clues lead us to propose a novel approach that considers the advantages of connected component analysis like Maximally Stable Extremal Region (MSER) properties, edge and gradient analysis like stroke width information, and texture analysis like the combination of Laplacian with wavelet sub-bands of text components, for text detection in video and natural scene images. This results in a hybrid approach for text detection of any orientation in this work. Therefore, the
contributions of the proposed approach are three folds: (1) A novel way to combine Laplacian with wavelet sub-bands to fuse images at different levels to identify candidate text pixels, which we called multi-spectral fusion, (2) The exploration of MSER along with the stroke width transform in a new way to identify candidate text regions, and (3) The alignment of character components using the nearest neighbor distance to restore full text lines of arbitrary orientations.

III. PROPOSED APPROACH

The proposed approach consists of four steps. In the first step, as we are inspired by the work presented in [6] for multi-oriented video text detection using the combination of Laplacian and Fourier, we propose a novel idea of convolving Laplacian with wavelet sub-bands at different levels in the frequency domain to enhance text pixels through a fusion concept, which results in multi-spectral fusion [29]. The fused images are subjected to fuzzy k-means clustering to classify the Candidate Text Pixels (CTP). Since video and natural images are complex and text has large variations in font, font size, color, orientation, etc., conventional methods such as the non-fuzzy k-means clustering, the Max-Min clustering method and the adaptive thresholding technique do not work well. The reason is that the conventional k-means clustering produces inconsistent results because of random guess selection, the Max-Min clustering method is not accurate enough due to the lack of discriminations between text and non-text values, and the adaptive thresholding technique does not give good results because it is hard to decide threshold values dynamically for different situations. Therefore, we prefer to use fuzzy K-means clustering, which classifies text and non-text pixels based on the probability of either text or non-text pixels with a membership function but not direct values. It is noticed that [6] Laplacian helps in distinguishing text pixels as it gives high positive and negative values for text pixels and low values for non-text pixels. In the same way, Fourier in the frequency domain provides high coefficients for text pixels and low coefficients for non-text pixels. Therefore, in this work, we propose to combine Laplacian with wavelet sub-bands for better enhancement because wavelet sub-bands have both low and high pass filters, while Fourier behaves either as a low pass or a high pass filter to eliminate noisy pixels in text detection. To tackle the problems of multi-size and multi-contrast texts, we combine Laplacian with wavelet sub-bands at different levels through fusion. This is illustrated in Fig. 1, where (a) is a sample text line image, (b) is the profile given by the Laplacian alone, (c) is the profile given by the Laplacian with Fourier, and (d) is the profile given by the Laplacian with wavelet. Fig. 1 shows that the profile given by Laplacian-wavelet has high sharp peaks for text pixels at stroke edges and low sharp peaks for non-text pixels occurring inside strokes, inside components and between components as compared with Laplacian alone and Laplacian-Fourier. Note that Fig. 1(d) shows the result of the average of Laplacian with high frequency sub-bands at the first level.

In the second step, for candidate text pixels, we explore Maximally Stable Extremal Regions (MSER) to group the candidate text pixels into text regions. MSER has been successfully used for classifying text and non-text components in the past [25, 30, 31]. In this work, we propose to modify MSER as sMSER along with Stroke Width Transform (SWT) [13] to cluster candidate text pixels as text regions, which we call Candidate Text Regions (CTR).

In the third step, we introduce Mutual Nearest Neighbor Clustering (MNN) for the CTR image to group candidate text regions that belong to the same text line. The proposed approach compares the geometrical properties of CTR before grouping them into a single one. The output of this step is said to be seed clusters that represent a text line. Sometimes, this step may eliminate text components due to mismatching. So in the fourth step, we present a new symmetry driven nearest neighbor process for each seed cluster to restore missing text clusters, which results in a full text line of any orientation. The symmetry is defined as the distance between two nearest neighboring components. The pipeline of the proposed approach can be seen in Fig. 2.

![Fig. 1: Laplacian values for text (edge) and non-text pixels](Image)
A. Multi-Spectral Fusion for Candidate Text Pixels Detection

As Fig. 1 illustrates, Laplacian-wavelet combination is good for finding fine details at edges in the image. It is also evident [7, 20] that Haar wavelet and its decomposition with moments have been deployed successfully for text detection in video. This motivates us to propose Haar wavelet for text detection in this work.

We use a two-level wavelet process to detect candidate text pixels (CTP) for the input image. First, we apply wavelet decomposition to obtain sub-bands, namely, LH, HL and HH for level-1 and level-2. Next, the proposed approach convolves Laplacian of different directions with the respective three sub-bands to enhance text pixel information in different directions. Then text pixels are separated from the three enhanced images with the help of fuzzy k-means clustering, which results in three text clusters. Candidate text pixels are obtained by performing an intersection operation on the three text clusters. This process is called intra fusion operation which outputs Candidate Text Pixels-1 (CT1). The same process is also used for wavelet decomposition level-2, which outputs Candidate Text Pixels-2 (CT2). Due to the intersection operation, it is possible to lose some text pixels. Therefore, we apply fuzzy k-means clustering with k=2 on the smoothed Laplacian of the input image without sub-bands, which outputs Candidate Text Pixels-0 (CT0). To obtain true Candidate Text Pixels (CTP), we perform a union operation on CT0, CT1 and CT2, which we call the inter fusion operation. In theory, the process can be extended to more levels of wavelet decomposition. However, according to our experiments, two levels are sufficient to achieve good results. The complete flow and steps are shown in Fig. 3.

An overview of wavelet analysis is as follows. Suppose \( \phi(x) \) is the scaling function of one-dimensional wavelet analysis, \( \psi(x) \) is the corresponding wavelet function, the relationship between two-dimensional scaling function and the 2D wavelet function can be defined by

\[
\phi(x, y) = \phi(x) \ast \phi(y) \\
\psi^1(x, y) = \phi(x) \ast \psi(y) \\
\psi^2(x, y) = \psi(x) \ast \phi(y) \\
\psi^3(x, y) = \psi(x) \ast \psi(y)
\]

where \( \phi(x, y) \) is the 2D scaling function, \( \psi^1, \psi^2, \psi^3 \) are the three 2D wavelet functions. The 2D multi-scale DWT can be defined as

\[
S_j f(m, n) = \int \int f(x, y) 2^j \phi(x - 2^j m, y - 2^j n) dx dy \\
W_j^1 f(m, n) = \int \int f(x, y) 2^j \psi_1^1(x - 2^j m, y - 2^j n) dx dy \tag{2} \\
W_j^1 f(m, n) = \int \int f(x, y) 2^j \psi_1^2(x - 2^j m, y - 2^j n) dx dy \\
W_j^1 f(m, n) = \int \int f(x, y) 2^j \psi_1^3(x - 2^j m, y - 2^j n) dx dy
\]
where \( f(x,y) \) is a gray image, \( S_j f(m,n) \) is the low-frequency sub-band of \( f(x,y) \) on scale \( j \), \( W_j^1 f(m,n) \), \( W_j^2 f(m,n) \), \( W_j^3 f(m,n) \) are the high-frequency sub-bands of \( f(x,y) \) in the horizontal, vertical and diagonal directions.

When \( x \leq c \) and \( x \geq c \), we can see non-text pixels are eliminated clustering on the three enhanced images can be respectively else it is the non-text one. The effect of fuzzy k-means considers the cluster that gives the higher mean as the text cluster, with \( k=2 \) because it is unsupervised and does not require any non-text pixels, we propose to use fuzzy k-means clustering with Laplacian masks of different directions namely horizontal, vertical and diagonal with sub-band images in the frequency domain, resulting in three enhanced spectral images. For instance, the horizontal Laplacian mask and its convolution with LH sub-band is as follows:

\[
\frac{\partial^2 f}{\partial x^2} = f(x+1,y) + f(x-1,y) - 2f(x,y) \quad (3)
\]

We define the mask as \( W^1 L ap \). In the same way, the proposed approach obtains spectral images for the other directions. The horizontal, vertical and diagonal masks of Laplacian are defined respectively as follows:

\[
W^2 L ap = f(x,y+1) + f(x,y-1) - 2f(x,y) \quad (4)
\]

\[
W^3 L ap = f(x+1,y) + f(x-1,y) + f(x,y+1) + f(x,y-1) - 4f(x,y)
\]

In a short form, we can define the above convolution operations as follows:

\[
\begin{align*}
W_j^1 L ap &= W_j^1 L ap \otimes W_j^1 f(m,n) \\
W_j^2 L ap &= W_j^2 L ap \otimes W_j^2 f(m,n) \\
W_j^3 L ap &= W_j^3 L ap \otimes W_j^3 f(m,n)
\end{align*}
\]

(5)

Note that \( \ast \ast \) denotes multiplication and \( \otimes \) indicates convolution in all the equations in this work.

The main reason to choose different directions of Laplacian masks to convolve with sub-bands is to handle arbitrary orientation of text in video. The inverse transform is used to convolve Laplacian masks of different directions namely horizontal, vertical and diagonal with sub-band images in the frequency domain, resulting in three enhanced spectral images. For instance, the horizontal Laplacian mask and its convolution with LH sub-band is as follows:

\[
\frac{\partial^2 f}{\partial x^2} = f(x+1,y) + f(x-1,y) - 2f(x,y) \quad (3)
\]

We define the mask as \( W^1 L ap \). In the same way, the proposed approach obtains spectral images for the other directions. The horizontal, vertical and diagonal masks of Laplacian are defined respectively as follows:

\[
W^2 L ap = f(x,y+1) + f(x,y-1) - 2f(x,y) \quad (4)
\]

\[
W^3 L ap = f(x+1,y) + f(x-1,y) + f(x,y+1) + f(x,y-1) - 4f(x,y)
\]

In a short form, we can define the above convolution operations as follows:

\[
\begin{align*}
W_j^1 L ap &= W_j^1 L ap \otimes W_j^1 f(m,n) \\
W_j^2 L ap &= W_j^2 L ap \otimes W_j^2 f(m,n) \\
W_j^3 L ap &= W_j^3 L ap \otimes W_j^3 f(m,n)
\end{align*}
\]

(5)

Note that \( \ast \ast \) indicates multiplication and \( \otimes \) indicates convolution in all the equations in this work.

The main reason to choose different directions of Laplacian masks to convolve with sub-bands is to handle arbitrary orientation of text in video. The inverse transform is used to convert frequency images to the respective spatial images for post processing. It is evident that the obtained three enhanced spectral images give high values for text and low values for non-text pixels because of the advantage of Laplacian and wavelet sub-bands combination. To separate text pixels from non-text pixels, we propose to use fuzzy k-means clustering with \( k=2 \) because it is unsupervised and does not require any threshold for classification. Since it is unsupervised, we consider the cluster that gives the higher mean as the text cluster, else it is the non-text one. The effect of fuzzy k-means clustering on the three enhanced images can be respectively seen in Fig. 5, where we can see non-text pixels are eliminated and text pixels are retained for all the three enhanced images. The steps of fuzzy k-means clustering are shown in Algorithm 1.

Algorithm 1 Fuzzy K-means clustering

**Input:** Peak values: \( \{x_1,x_2,...,x_{m*n}\} \) (m*n is the size of LOG image, \( x_i \) is the peak value of i-th pixel);

\( K \) (number of peak levels).

**Output:** \( C \) \{set of generated centroids of K class\}

\( U \) (probability matrix)

1: **Initial the K membership function:**

\( c_0 = \min\{x_1,x_2,...,x_{m*n}\} \), \( c_{K+1} = \max\{x_1,x_2,...,x_{m*n}\} \)

\( c_1, c_2, ..., c_K \) as:

\[
C_j = c_0 + \frac{(x_j - c_{j+1})}{c_{j+1} - c_j}
\]

(\( j = 1,2,...,K \)) using the following equations:

\[
C_j = \frac{\sum_i u_{ij} x_i}{\sum_i u_{ij}}
\]

2: Set \( U=0 \) (U is the probability matrix), \( U_{ij} \) denotes the peak value of i-th pixel in LOG image belongs to j-th level.

\[
\begin{align*}
U_{il} &= 1, U_{ij+1} = 0 & \text{when} \ x_i \leq c_1 \\
U_{lK} &= 1, U_{ij+K} = 0 & \text{when} \ x_i > c_K \\
U_{ij} &= \frac{c_{j+1} - x_i}{c_{j+1} - c_j} U_{ij+1} + 1 - U_{ij+1} & \text{when} \ c_j < x_i \leq c_{j+1}
\end{align*}
\]

3: Update the class centroids \( C_j \) (\( j = 1,...,K \)) using the following equ:

\[
C_j = \frac{\sum_i u_{ij} x_i}{\sum_i u_{ij}}
\]

4: Repeat step 2 and step 3 until the \( C_j \) (\( j = 1,...,K \)) are unchanged.

In Algorithm 1, step 1 sets two membership functions as an initial fuzzy partition. In step 2, each element in the probability matrix \( U \) is computed according to the three rules listed. Step 3 updates class centroids to reflect new data distribution. The algorithm will be terminated if the class centroids are unchanged, otherwise, step 2 and step 3 are repeated. The final probability matrix \( U \) and class centroids \( C \) will be generated by this way. For each sub-band, we select the pixels with the highest probability value of the high peak level in \( U \) as text pixels.

We perform morphological operations to group the pixels that have close proximity. It is observed from Fig. 5 that the enhancement process retains most of the text pixels in the three respective sub-band images along with a few non-text pixels. In order to eliminate non-text pixels and retain text pixels, we propose to fuse the three images by performing an intersection operation, which gives one fused image containing most significant text pixels as shown in Fig. 6(a), where one can see most of the non-text pixels are removed. We denote the fused result as CT1 because the proposed approach fuses...
Laplacian-Sub-bands at the first level of wavelet decomposition. The fusion is considered as intra fusion since it fuses the sub-bands of the same level. The reason to perform intersection operation for fusion is that text detection does not require full text information. If we get at least one component for one full text line as a representative, it is enough to restore the full text line using restoration approach (this will be discussed in Section D). As a result, the fusion operation gives representatives that exist in all the three images. The advantage of this step is that it eliminates most of the non-text pixels as shown in Fig. 6(a).

![Fig. 6: The effect of intra fusion](image)

It is observed that scene texts in video and natural scene images generally have large variations in font size. Therefore, we propose to deploy the proposed approach at multi-level wavelet decomposition. The proposed approach repeats all the steps from the input to the fused image for the second level as shown in Fig. 6(b), where we can see some of the non-text pixels are removed, which are denoted by CT2. The method preforms a union operation for CT1 and CT2 to combine significant text pixels as shown in Fig. 6(c), where we can see a few missing text pixels at the first level are restored along with non-text pixels. In this work, we stop at the second level because our experimental results show that two levels are enough on the dataset. In the same way, one more example can be seen in Fig. 7(a), where one can see texts of different font sizes. For the image in Fig. 7(a), the results at the first level miss the text “e-Card” due to font size variations compared to other texts in CT1 as shown in Fig. 7(b). For the same image, the results at the second level restore the text “e-Card” in CT2 as shown in Fig. 7(c). In this way, the proposed approach obtains intra fused images of the first and the second levels for every input image, denoted by CT1 and CT2, respectively.

![Fig. 7: Text pixels restoration using multi-level fusion](image)

It is noticed from Fig. 7(b) and (c) that both the results contain significant text information. Rather than performing the intersection operation as it is done for intra fusion, we propose to perform a union operation for fusing CT1 and CT2 images, which gives complete text information as shown in Fig. 7(d), where we can see almost all the text information. This output is called inter fusion operation as it combines two intra fused results obtained by the respective levels. However, we can still see that some text pixels are missing, which have too low contrast compared to the input image in Fig. 7(a). This is possible because of the intersection operation during intra fusion. To restore such low contrast text information, we propose to apply fuzzy k-means clustering with k=2 on the smoothed Laplacian of the input frame without wavelet sub-bands convolution to classify text pixels from non-text pixels, which we call CT0. This is shown in Fig. 8(a) and Fig. 8(b), where it is noticed that low contrast text pixels are classified as text pixels along with non-text pixels. To extract such Candidate Text Pixels (CTP) in CT0, we propose a smoothing operation by moving a window over CT0. In other words, for each text pixel candidate in the inter fused image, we perform a morphological operation over CT0 to restore missing text pixels. The smoothing operation is defined as

\[
CTP = (\text{Morph}_{\text{wind}} \& CT0) | \text{Inter}_{\text{fusion}}
\]

\[\forall t \in (-N/2, N/2) \text{ s.t. } p(x - t, y - t) = 1\]

where \((x,y)\) are all the text pixel candidates and \(N\) is the width of the morphological window on the inter fusion image.

This operation for both example images in Fig. 5(a) and Fig. 7(a) is respectively illustrated in Fig. 8(c) and Fig. 8(d), where we can notice how low contrast pixels are restored. The output of this operation is called Candidate Text Pixels (CTP).

![Fig. 8: Candidate text pixels using multi-spectral fusion approach](image)

Note that from literature [6] and the illustration shown in Fig. 1, the Laplacian operation provides high positive and negative peaks at edge pixels when there is a transition from background to foreground or vice versa (zero crossing) since it is a second order derivative. This is the main cue to identify the presence of text. As a result, Laplacian operation helps in enhancing text pixel information in video images. However, due to complex background, the Laplacian operation generates noisy pixels during operation. Inspired by the work proposed in [6] for enhancing text information by combining Laplacian with Fourier, we propose to combine Laplacian with Wavelet to enhance low contrast text pixels in video images. In [6], Fourier has been used as an ideal low pass filter to remove noisy pixels generated by the Laplacian operation for enhancement, while in this work we propose Wavelet decomposition at different levels to extract spatially coherent horizontal, vertical and diagonal edge pixels, which represent true text pixels despite noisy ones produced by the Laplacian operation. This is because the pixels in texts distribute in horizontal, vertical and diagonal directions with spatial coherence prominently. This is evident from [7, 20] where a combination of wavelets and moments is proposed for text detection in video by exploiting wavelet decomposition and spatial information given by moments. This is also illustrated in Fig. 9, where we can notice from the results given by Laplacian alone in Fig. 9(a) that it enhances both text and non-text pixels with noises, while from the results given by
Wavelet alone in Fig. 9(b), we can notice that it sharpens text pixels with neighboring pixels compared with non-text pixels and the Laplacian result. On the other hand, Fig. 9(c) shows that the combination eliminates most of the non-text pixels and retains text pixels as this combination chooses true text pixels. Since our objective is to achieve good results for both video and natural scene images where images generally can have complex background, the combination of Laplacian and Wavelet is thus necessary for this target.

(a) Laplacian (b) Wavelet sub-bands (c) Laplacian+Wavelet
Fig. 9: Candidate text pixels for only Laplacian, Wavelet and Laplacian+Wavelet

B. sMSER for Text Candidate Detection
It is found from the CTP of the input frame given by the previous section that the CTP still contains some non-text candidate pixels due to background complexity. To eliminate such non-text candidate text pixels, we propose to explore MSER along with the stroke width property because it is true that MSER is a successful approach for studying the characteristics of connected components [25, 30, 31], especially for text detection and recognition in video and natural scene images. Most of the approaches in the literature apply MSER for the whole image to study the characteristics of text components. However, the use of MSER for the whole image is not advisable because it sometimes largely misclassifies non-text components as text ones due to complex background and low resolution. This may lead to poor performance of this approach. To avoid this problem, we apply MSER for only the candidate text pixels given by the approach presented in Section III.A. Since CTP is a binary image and MSER requires gray information, the proposed approach extracts gray values in the input image corresponding to the candidate text pixels in the CTP image. Even then, conventional MSER still gives poor results because of background color variations within character components. In view of this, MSER detects multiple overlapping sub-parts of the same components instead of a single component. Therefore, we perform a mean filter operation over the gray values of the input image that correspond to the edges of the CTP in the CTP image. This operation smoothens edges irrespective of multiple colors in a single component. In other words, it is a simple averaging filter, which performs 3×3 mask operations for every pixel in the CTP image. The output of smoothing is considered as the input for MSER, which we name sMSER in this work. It is evident from the illustration shown in Fig. 10 that the MSER proposed in [30] gives 53 components for 8 character components, while the proposed sMSER gives 39 components. Besides, the output of sMSER is smoother as we can notice the boundaries of the components in Fig. 10(b) compared to the results in Fig. 10(a). As a result, the proposed sMSER helps in preserving the shapes of the characters as shown in Fig. 10(b), which we call text candidates. This is the advantage of the proposed sMSER.

(a) Existing MSER (b) Proposed sMSER
Fig. 10: Preserving shape of the characters by sMSER

However, the proposed sMSER does not remove non-text candidate text pixels from the CTP image. Rather, it helps in preserving the shapes of the characters. This motivates us to use stroke width features for removing non-text candidate pixels because it is a fact that characters generally have constant stroke width throughout each character [8, 13]. The stroke width is estimated as follows [13]: For every candidate text pixel occurring at the edge of a stroke, the method finds its gradient direction which is perpendicular to the stroke direction and then moves in the gradient direction until it reaches another candidate text pixel in the edge image. The distance between the starting pixel and the ending pixel is considered as the stroke width distance. In this way, for each text candidate, we calculate the mean and the standard deviation of stroke width distances at the component level to eliminate non-text candidates. Then we set a condition to remove non-text candidates as std/mean>swt_0. Similarly, we also propose one more rule based on the density of pixels in text candidates as an occupancy ratio, which is defined as

\[ \text{occu ratio} = \frac{\text{Area(Region)}}{\text{Area(Bounding box)}} \]

if \(\text{occu ratio} > t_{o}\) then non-text

(a) sMSER (b) SWT (c) Occupation ratio
Fig. 11: Non-text candidate elimination

If a text candidate satisfies the above condition, we consider it as a non-text component and hence we eliminate it. The thresholds mentioned above are determined experimentally. The effect of sMSER and the rules for the full image are shown in Fig. 11, where (a) is the result of sMSER, (b) is the result of rule based on stroke width distances, and (c) is the result of occupancy ratio. It is noticed from Fig. 11(b) and (c) that the rules eliminate a few non-text candidates successfully without affecting text candidates.

C. Mutual Nearest Neighbor Clustering for Seed Cluster Detection
Fig. 11 shows that due to arbitrary orientation and complex background, it is hard to remove non-text candidates completely. Therefore, to reduce the effect of non-text candidates, we propose Mutual Nearest Neighbor (MNN) clustering for grouping the text candidates that share common properties [32]. It is noted that character components in a text line usually share uniform color, size, and distance. The MNN is defined as the following. Let A and B be the two nearest neighbors of a text line. If A satisfies a particular property such as size, then B should satisfy the same property as A. Thus we consider A and B as mutually nearest neighbors and hence we classify them as one cluster. In this way, the proposed MNN...
gives different clusters for the text candidates given by Section III.B as it is defined in the following equation:

\[ M_k(x) = \{ x_i \in D \mid x_j \in N_k(x) \land x_i \in N_k(x_j) \} \]  

where \( D \) is the input data consisting of text candidates \((x_i)\), \( N_k(x) \) is a set of \( k \) nearest neighbors of \( x \). Here, we consider \( k=2 \) because every text candidate in a text line has at most two nearest neighbors. The proposed MNN considers two properties, namely distance to find the nearest neighbor and aspect ratio, which is the ratio of width to height to find the similarity between text candidates. This is illustrated in Fig. 12, where we can see three clusters shown respectively in Fig. 12(a)-(c) given by MNN for the text candidates in Fig. 11(c). It can be seen that some of the non-text candidates are eliminated while other non-text candidates are clustered as text candidates.

![Cluster-1](a), (b) Cluster-2 (c), and (d) Seed cluster (d)

Fig. 12: MNN for grouping similar text candidates to find seed text cluster

It is observed that the stroke width distances of the text components in the clusters appear to be almost similar. Therefore, we test the stroke width consistency of the text candidates in each cluster to classify it as either a text cluster or a non-text cluster as shown in Fig. 12(d), where we can notice that non-text clusters are removed. Here, we calculate stroke width as described in Section III.B for each text candidate in the cluster, and test the mean of stroke widths of the maps of different text candidates in the cluster to verify whether the text candidates exhibit consistent mean stroke widths. This results in seed text clusters, which can be used for traversing text lines with the help of symmetry driven growing method.

D. Symmetry Driven Nearest Neighbor for Arbitrary Text Line Detection

For each text candidate, say A in a seed cluster, the proposed approach finds the nearest neighbor by using Euclidean distance between the centroids of the text candidates, say B. The same distance is calculated for B to find the nearest neighbor. If B finds A as the nearest neighbor then both A and B are said to have satisfied the symmetry property. This process continues until no text candidates exist in the seed cluster. There are other approaches [7, 26, 33] that explore the growing process in different ways to extract arbitrarily oriented text lines, but these approaches do not use the symmetry concept. Thus the main advantage of this process is that it works for any direction of text line. In addition, the distance between text candidates can be used for word segmentation. This process works well because of the fact that the space between characters is usually less than the space between words and text lines. The distance between text candidates in the seed cluster is calculated as follows:

\[ \min D_u < \max D_g \]

\[ \max D_u > \min D_g \]

\[ 1/2 \text{mean}(D_g) < \text{mean}(D_u) < 2 \text{mean}(D_g) \]

where \( D_u, D_g \) are the distances between the text candidates in the seed cluster. This process gives a pair of text candidates as shown in Fig. 13(a), where we can see the first two text candidates (F and U) are merged together as one component. Fig. 13(b) shows the next pair of text candidates, namely, U and N. Therefore, all the three text candidates (F, U and N) are merged and fixed in one bounding box that snugly encloses the word “FUN”. In this way, the symmetry driven nearest neighbor process fixes a flexible bounding box that encloses any arbitrary text line as shown in Fig. 13(c). The final text line extraction can be seen in Fig. 13(d). Another advantage of this symmetry driven nearest neighbor process is that suppose a seed cluster contains only two text candidates for the whole text line, when we apply this process on the edge image of the input frame by considering the text candidate as a seed text candidate, we can restore the full text lines with the help of seed text candidates. Therefore, we can conclude that if MNN clustering misses some text candidates, the symmetry nearest neighbor growing process will restore the missing text candidates with the help of the edge image of the input frame.

![Pair](a), (b) Word (c) All pairs (d) Full text line

Fig. 13: Symmetry driven growing for arbitrary text line detection

In this work, we use Stroke Width Transform (SWT) [13] as a supporting feature since it is a commonly used baseline approach for text detection in natural scene images reported in the literature. Fig. 14 shows the superior performance of the proposed approach as compared to the poor results of the baseline. Since SWT works at edge component levels, it expects good shapes of characters in order to classify text components properly from non-text components. However, since video contains both low and high contrast texts with complex background, it is hard to preserve the shapes of characters. Therefore, SWT gives poor results for video frames. On the other hand, the proposed approach is developed for tackling both low and high contrast texts in video as well as natural scene images. The proposed idea of combining Laplacian with wavelet decomposition and sMSER contributes more for finding accurate text candidates irrespective of the image type.
from ICDAR 2013 and 300 from MSRA for experimentation in this work. In total, 2019 images are used, of which 1257 are video frames and 762 are natural scene images. We believe this dataset helps us to show the generic nature of the proposed approach.

Since there is no ground truth for our video data and Hua’s data, we count manually the number of actual text blocks for calculating recall, precision and f-measure. We follow the instructions and the evaluation scheme as proposed in ICDAR 2011 for calculations. In addition, for our dataset, we count recall, precision and f-measure at text line level as well as at word level as done in the following works: [6, 7, 8, 10, 12, 20, 21, 26]. For ICDAR 2013 video data, since the ground truth at the word level is available, we follow the evaluation scheme as in ICDAR 2011. In the same way, for natural scene image data of ICDAR 2011 and ICDAR 2013, we use the same evaluation scheme as in ICDAR 2011 reading competition [36]. However, for MSRA data, since the ground truth is available at text line level, we use the same evaluation scheme as in ICDAR 2011 at text line level instead of word level. The precision, recall and f-measure can be defined mathematically as

\[
p = \frac{\sum_{i \in \text{R}} m(r, T)}{|E|} \quad \text{and} \quad r = \frac{\sum_{i \in \text{E}} m(r, E)}{|T|}
\]

where \(m(r, R)\) is the best match for a block \(r\) in a set of block \(R, E\) and \(T\) are our estimated block and the ground truth block, respectively. The \(f\) measure is defined using recall, precision as

\[
f = \frac{1}{(\alpha / p + \alpha / r)} \quad \text{where} \quad \alpha = 0.5 \quad \text{for all the experiments in this work.}
\]

In addition to recall, precision and F-measure, we also evaluate the performance of the proposed approach in terms of efficiency on the video data. For this, we calculate the Average Processing Time (APT) for the dataset.

We compare the proposed approach with the following baseline approaches to show the effectiveness of the proposed approach. First, we implement the approach based on stroke width transform [30], which explores modified stroke width distances of character components for text detection in natural scene images and is considered as the state of the art approach. Next, we implement the approach based on Laplacian [6], which explores the combination of Laplacian and Fourier for text detection in video frames. This is considered as the state of the art methods for video text detection. Additionally, another latest approach based on characterness [16] is implemented, which explores the saliency of objectness, Bayesian classifier and Markov random field for text detection in natural scene images. We run these three approaches on all the datasets for comparative study.

**B. Analyzing Steps of the Proposed Approach**

To validate the strength of each step, we conduct an experiment on 100 sample images chosen randomly from all the databases to evaluate each step individually in terms of recall, precision and f-measure. The 100 sample images comprise 44% from the video data and the remaining 56% from the natural image data. The results are listed in Table I, which shows that all the steps contribute to the performance of the proposed approach. If the proposed method misses any one of the step, it affects the overall performance. Therefore, to achieve the best accuracy
compared to the state of the art approaches, small incremental improvements also do make noticeable differences.

Table I: Analyzing contributions of each step of the proposed method

<table>
<thead>
<tr>
<th>Steps of the Proposed Method</th>
<th>P</th>
<th>R</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laplacian alone</td>
<td>0.68</td>
<td>0.65</td>
<td>0.66</td>
</tr>
<tr>
<td>Wavelet sub-bands alone</td>
<td>0.72</td>
<td>0.66</td>
<td>0.68</td>
</tr>
<tr>
<td>Without smoothing in III.B</td>
<td>0.73</td>
<td>0.68</td>
<td>0.70</td>
</tr>
<tr>
<td>Without Stroke width in III.B</td>
<td>0.56</td>
<td>0.64</td>
<td>0.60</td>
</tr>
<tr>
<td>Without Stroke width consistency in III.C</td>
<td>0.62</td>
<td>0.65</td>
<td>0.63</td>
</tr>
<tr>
<td>Without symmetry checking in III.D</td>
<td>0.76</td>
<td>0.67</td>
<td>0.71</td>
</tr>
<tr>
<td>Proposed Laplacian-Wavelet</td>
<td>0.76</td>
<td>0.69</td>
<td>0.72</td>
</tr>
</tbody>
</table>

Since the combination of Laplacian-Fourier presented in [6] for text detection looks similar to the proposed approach, we conduct another experiment on the same 100 images to show how the proposed combination, namely, Laplacian-Wavelet, is different in terms of recall, precision and f-measure from Laplacian-Fourier in Table II. Though the proposed work is inspired by the work in [6], both the objective and the concepts differ significantly. The main difference is that Laplacian-Fourier is proposed for video text detection but not both video as well as natural scene text detection in the proposed approach. Further, in [6], both Laplacian and Fourier are complementing each other because Laplacian is used to enhance text information while Fourier is used as an ideal low pass filter to remove the noises generated by the Laplacian operation. In the proposed work, we propose the Laplacian-Wavelet combination to select true text pixels. In addition, we exploit the advantages of wavelet decomposition, which helps in extracting spatial coherence of text pixels in different directions and multi-sized fonts for text detection. Therefore, the proposed approach using Laplacian-Wavelet gives better results than the proposed approach with Laplacian-Fourier as reported in Table II.

Table II: Performance of the proposed method Laplacian-Wavelet vs Laplacian-Fourier

<table>
<thead>
<tr>
<th>Methods</th>
<th>P</th>
<th>R</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed method with Laplacian-Fourier</td>
<td>0.72</td>
<td>0.67</td>
<td>0.69</td>
</tr>
<tr>
<td>Proposed method with Laplacian-Wavelet</td>
<td>0.76</td>
<td>0.69</td>
<td>0.72</td>
</tr>
</tbody>
</table>

C. Experiments on Video Data

Sample qualitative results of the proposed and the existing methods for our video data are shown in Fig. 15, where for the input frames of different backgrounds with scene texts in (a), the proposed approach detects almost all the texts successfully, while Chen et al.’s method [30] misses some text components for all the three frames due to rigid conditions used for false positive elimination. Shivakumara et al.’s method [6] is good for all the three frames except false positives, and Li et al.’s approach [16] also detects almost all the text lines in the input frames. However, Li et al.’s method [16] misses the text components for the second frame. The quantitative results of the proposed and the existing methods at the text line level and the word level are reported in Table III and Table IV, respectively. It is observed from Table III and Table IV that the proposed approach is better at recall and f-measure than the existing approaches for both text and word levels. Li et al.’s method is the best at recall for both text line and word levels in text detection. The main reason for the poor accuracies of the existing approaches is that the approaches depend more on the connected component analysis and the shape of the characters. Due to low resolution and complex background of video, it is hard to preserve the shape of a character. On the other hand, the proposed approach is able to preserve fine details of character components through the enhancement by Laplacian-Wavelet combination and sMSER. Hence, it gives good results compared to the existing approaches. When we compare Table III and Table IV, the accuracy at the word level is a little lower than that at the text line level. This is because of the errors in the segmentation of words from text lines especially when a text line is curved.

Table III: Performance on our own video data at text line level

<table>
<thead>
<tr>
<th>Method</th>
<th>precision</th>
<th>recall</th>
<th>f-measure</th>
<th>APT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>0.72</td>
<td>0.67</td>
<td>0.69</td>
<td>14.37</td>
</tr>
<tr>
<td>Chen et al.[30]</td>
<td>0.52</td>
<td>0.43</td>
<td>0.47</td>
<td>10.45</td>
</tr>
<tr>
<td>Shivakumara et al. [6]</td>
<td>0.58</td>
<td>0.61</td>
<td>0.59</td>
<td>5.12</td>
</tr>
<tr>
<td>Li et al. [16]</td>
<td>0.73</td>
<td>0.62</td>
<td>0.67</td>
<td>17.52</td>
</tr>
</tbody>
</table>

Table IV: Performance on our own video data at Word Level

<table>
<thead>
<tr>
<th>Method</th>
<th>precision</th>
<th>recall</th>
<th>f-measure</th>
<th>APT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>0.71</td>
<td>0.66</td>
<td>0.68</td>
<td>13.87</td>
</tr>
<tr>
<td>Chen et al.[30]</td>
<td>0.59</td>
<td>0.44</td>
<td>0.50</td>
<td>9.84</td>
</tr>
<tr>
<td>Shivakumara et al. [6]</td>
<td>0.54</td>
<td>0.62</td>
<td>0.58</td>
<td>5.12</td>
</tr>
<tr>
<td>Li et al. [16]</td>
<td>0.72</td>
<td>0.60</td>
<td>0.66</td>
<td>16.87</td>
</tr>
</tbody>
</table>

The qualitative results of the proposed and the existing approaches for ICDAR 2013 video data are shown in Fig. 16,
where the proposed approach gives good results, while Chen et al.’s and Li et al.’s methods miss some text components and Shivakumara et al.’s method gives some false positives. Table V shows that Li et al.’s method is the best at precision as it does not give false positives compared to the proposed approach and the other existing approaches. However, Li et al.’s method is low at recall compared to Shivakumara et al.’s method and the proposed approach. Therefore, the proposed approach outperforms the existing approaches in terms of recall and f-measure.

Table V. Performance on ICDAR 2013 Video at Word level

<table>
<thead>
<tr>
<th>Method</th>
<th>precision</th>
<th>Recall</th>
<th>f-measure</th>
<th>APT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>0.76</td>
<td>0.66</td>
<td>0.70</td>
<td>16.29</td>
</tr>
<tr>
<td>Chen et al. [30]</td>
<td>0.54</td>
<td>0.42</td>
<td>0.47</td>
<td>11.57</td>
</tr>
<tr>
<td>Shivakumara et al. [6]</td>
<td>0.58</td>
<td>0.66</td>
<td>0.62</td>
<td>6.85</td>
</tr>
<tr>
<td>Li et al. [16]</td>
<td><strong>0.78</strong></td>
<td>0.63</td>
<td>0.70</td>
<td>20.03</td>
</tr>
</tbody>
</table>

We also show some sample results of the proposed and the existing approaches for the benchmark dataset [35] in Fig. 17, in which we can see that the proposed approach gives good results for the input frame, while the existing approaches either miss some text components or give false positives. Table VI shows that the proposed approach gives the best accuracy over all the existing approaches. When we compare Table III, Table IV, Table V and Table VI, the f-measure for Hua’s data is higher than other video data because the accuracy is counted at the text line level and the dataset is small without arbitrary orientations.

Table VI. Performance on Hua’s data at text line level

<table>
<thead>
<tr>
<th>Method</th>
<th>precision</th>
<th>recall</th>
<th>f-measure</th>
<th>APT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td><strong>0.75</strong></td>
<td>0.67</td>
<td><strong>0.71</strong></td>
<td>12.62</td>
</tr>
<tr>
<td>Chen et al. [30]</td>
<td>0.62</td>
<td>0.56</td>
<td>0.59</td>
<td>9.67</td>
</tr>
<tr>
<td>Shivakumara et al. [6]</td>
<td>0.64</td>
<td>0.65</td>
<td>0.65</td>
<td></td>
</tr>
</tbody>
</table>

It is observed from Table III to Table VI that the proposed approach is not as efficient as the approach in [6]. However, the proposed approach is better than the approach in [16] for all the experiments. While we aim at achieving good results for both video and natural scene images in the present work, its efficiency can be further improved in our future investigation. It is noted that the multi-spectral fusion step consumes more time compared to the other steps of the proposed approach. Therefore, we plan to optimize this step in future by finding a trade-off between multi-levels of sub-bands and the complexity of the problem. It is also noted that fuzzy K-means clustering also requires more time as it is an iterative method. In order to reduce the number of iterations, we plan to modify the clustering algorithm such that it converges more quickly. In the same way, symmetry driven growing takes more time as it involves the growth of text candidates. To reduce the growth duration, we would introduce linearity and non-linearity between text candidates rather than growing the whole text candidates.

D. Experiments on Natural Scene Data

Fig. 18 shows the qualitative results of the proposed and the existing approaches for complex background input images. It is observed from Fig. 18 that the proposed approach gives better results compared to the existing approaches. All the approaches detect texts well but the existing approaches give more false positives due to complex background. As a result, the accuracy of the existing approaches is lower than the proposed approach especially in terms of recall and f-measure for the ICDAR 2013 dataset according to the results reported in Table VIII. However, Table VII shows that the proposed approach is not the best when compared to methods by Yin et al. [15] and Yi and Tian [9]. The proposed approach is the second best in terms of recall and f-measure. The reason is that ICDAR 2011 data do not contain too many oriented text lines compared to ICDAR 2013 dataset. Since the proposed approach is developed to handle arbitrary orientations as well as video, it slightly loses accuracy as a trade-off for greater generality.
The proposed approach introduces mutual nearest neighbor clustering based on geometrical properties of text candidates to group text candidates of respective text lines into clusters. The symmetry driven growing process is proposed to extract arbitrary text lines based on the distance between text candidates in each cluster. According to our knowledge, this is the first attempt to detect text in both video frames and natural scene images with good accuracies. However, according to the results, the accuracy is still lower than that in document analysis, where usually the value would be more than 80%. We are planning to improve the accuracy of the current approach by making use of temporal frames in video through tracking in our future work.

Though our focus of the paper is to detect low contrast and low resolution texts, the proposed approach sometimes fails to detect texts properly if the image contains too small fonts, low resolution and too complex background as shown in Fig. 20. This is because the proposed approach targets both video as well as natural scene images with good results at the expense of too low resolution texts and small fonts. Therefore, there is a scope for improvement in future.

Table IX. Performance on MSRA-TD500 data at text line level

<table>
<thead>
<tr>
<th>Method</th>
<th>precision</th>
<th>recall</th>
<th>f-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>0.74</td>
<td>0.66</td>
<td>0.70</td>
</tr>
<tr>
<td>Chen et al. [30]</td>
<td>0.65</td>
<td>0.58</td>
<td>0.66</td>
</tr>
<tr>
<td>Shivakumara et al. [6]</td>
<td>0.60</td>
<td>0.62</td>
<td>0.62</td>
</tr>
<tr>
<td>Li et al. [16]</td>
<td>0.74</td>
<td>0.61</td>
<td>0.67</td>
</tr>
<tr>
<td>Yao et al. [5]</td>
<td>0.64</td>
<td>0.62</td>
<td>0.61</td>
</tr>
<tr>
<td>Kang et al. [39]</td>
<td>0.71</td>
<td>0.62</td>
<td>0.66</td>
</tr>
<tr>
<td>Yin et al. [15]</td>
<td>0.71</td>
<td>0.61</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Table VIII. Performance on ICDAR 2013 data word level

<table>
<thead>
<tr>
<th>Method</th>
<th>precision</th>
<th>recall</th>
<th>f-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chen et al. [30]</td>
<td>0.68</td>
<td>0.58</td>
<td>0.63</td>
</tr>
<tr>
<td>Shivakumara et al. [6]</td>
<td>0.62</td>
<td>0.64</td>
<td>0.63</td>
</tr>
<tr>
<td>Li et al. [16]</td>
<td>0.79</td>
<td>0.63</td>
<td>0.70</td>
</tr>
<tr>
<td>Neumann and Matas [37]</td>
<td>0.73</td>
<td>0.64</td>
<td>0.68</td>
</tr>
<tr>
<td>Liu et al. [38]</td>
<td>0.70</td>
<td>0.59</td>
<td>0.64</td>
</tr>
<tr>
<td>Epishtein et al. [13]</td>
<td>0.73</td>
<td>0.60</td>
<td>0.66</td>
</tr>
<tr>
<td>Epshtein et al. [13]</td>
<td>0.73</td>
<td>0.60</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Table VII. Performance on MSRA-TD500 data at word level

<table>
<thead>
<tr>
<th>Method</th>
<th>precision</th>
<th>recall</th>
<th>f-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chen et al. [30]</td>
<td>0.65</td>
<td>0.58</td>
<td>0.66</td>
</tr>
<tr>
<td>Shivakumara et al. [6]</td>
<td>0.60</td>
<td>0.62</td>
<td>0.62</td>
</tr>
<tr>
<td>Li et al. [16]</td>
<td>0.74</td>
<td>0.61</td>
<td>0.67</td>
</tr>
<tr>
<td>Yao et al. [5]</td>
<td>0.64</td>
<td>0.62</td>
<td>0.61</td>
</tr>
<tr>
<td>Kang et al. [39]</td>
<td>0.71</td>
<td>0.62</td>
<td>0.66</td>
</tr>
<tr>
<td>Yin et al. [15]</td>
<td>0.71</td>
<td>0.61</td>
<td>0.66</td>
</tr>
<tr>
<td>Risnumawan et al. [33]</td>
<td>0.70</td>
<td>0.68</td>
<td>0.69</td>
</tr>
<tr>
<td>Neumann and Matas [37]</td>
<td>0.73</td>
<td>0.64</td>
<td>0.68</td>
</tr>
<tr>
<td>Liu et al. [38]</td>
<td>0.70</td>
<td>0.59</td>
<td>0.64</td>
</tr>
<tr>
<td>Epishtein et al. [13]</td>
<td>0.73</td>
<td>0.60</td>
<td>0.66</td>
</tr>
</tbody>
</table>

REFERENCES


Fig. 19 shows some sample qualitative results of the proposed and the existing approaches for an input image containing multi-oriented text lines in both Chinese and English. It is noticed from Fig. 19 that the proposed approach detects both the text lines but misses two isolated Chinese characters. Therefore, the recall is lower than the precision for the proposed approach according to the results reported in Table IX. However, overall, the proposed approach outperforms the existing approaches in terms of recall, precision and f-measure. In summary, the main reason that leads to poor accuracies for the existing methods is that most of the existing approaches rely on training samples to train the classifier for classifying text and non-text. But, it is hard to collect training samples that represent non-text because there is no boundary for defining non-text. In addition, the use of classifiers limits their scope to a particular language. On the other hand, the proposed approach does not involve much tuning and does not require a large number of training samples compared to the existing approaches. The superiority of the proposed approach is seen in its ability in handling different situations, such as video, natural scene images and multi-script or arbitrary orientation ability.

Fig. 19: Experiments on MSRA natural scene data

(a) Input frame (b) Chen et al. (c) Shivakumara et al.
(d) Li et al. (e) Proposed approach

Fig. 20: Limitations of the proposed method

CONCLUSION AND FUTURE WORK

In this paper, we have proposed a novel idea of combining Laplacian with wavelet frequency sub-bands through fusion at multi-level to identify text candidates. We have explored Maximally Stable Extremal Regions along with stroke width distances for preserving fine details of text candidates. The proposed approach introduces mutual nearest neighbor clustering based on geometrical properties of text candidates to group text candidates of respective text lines into clusters. The symmetry driven growing process is proposed to extract arbitrary text lines based on the distance between text candidates in each cluster. According to our knowledge, this is the first attempt to detect text in both video frames and natural scene images with good accuracies. However, according to the results, the accuracy is still lower than that in document analysis, where usually the value would be more than 80%. We are planning to improve the accuracy of the current approach by making use of temporal frames in video through tracking in our future work.


Guozhu Liang is now a research graduate student at the Department of Computer Science and Technology, Nanjing University. His current interests are in the areas of image processing, computer vision and pattern recognition algorithms.

Palahlamkote Shivakumara is a Senior Lecturer in Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia. Previously, he was with the Department of Computer Science, School of Computing, National University of Singapore from 2008-2013 as a Research Fellow on Video text extraction and recognition project. He received B.S., M.Sc., M.Sc Technology by research and Ph.D degrees in computer science respectively in 1995, 1999, 2001 and 2005 from University of Mybose, Mybose, Karnataka, India.

From 1999 to 2005, he was Project Associate in the Department of Studies in Computer Science, University of Mybose. He worked as a Research Fellow in the field of image processing and multimedia in the Department of Computer Science, School of Computing, National University of Singapore from 2005-2007. He also worked as a Research Consultant in Nanyang Technological University, Singapore for a period of 6 months on image classification in 2007. He has published more than 150 research papers in national, international conferences and journals. He has been reviewer for several conferences and journals. He has been serving as Associate Editor for ACM Transactions Asian Language Information Processing (TALIP). His research interests are in the area of image processing, pattern recognition, including video text analysis.

Tong Lu received the PhD degree in computer science from Nanjing University in 2005. He received his M.Sc. and B.Sc. degree from the same university in 2002 and 1993, respectively. He served as Associate Professor and Assistant Professor in the Department of Computer Science and Technology at Nanjing University from 2007 and 2005. He is now a full Professor at the same university. He also has served as Visiting Scholar at National University of Singapore and Department of Computer Science and Engineering, Hong Kong University of Science and Technology, respectively. He also is a member of the National Key Laboratory of Novel Software Technology in China. He has published over 60 papers and authored 2 books in his area of interest, and issued more than 20 international or Chinese invention patents. His current interests are in the areas of multimedia, computer vision and pattern recognition algorithms/systems. Dr. Tong Lu was a member of ACM, IAPR, ISAI and a senior member of China Computer Federation (CCF). He is the Youth Associate Editor of Journal on Frontiers of Computer Science (FCS), and has served as the Secretary-general of CAD&CG Committee of Jiangsu Computer Federation in China since 2008. He has been member of the program committee or session chair of more than 10 international scientific conferences, and the Chair of Organization Committee of Youth Scholar Forum of State Key Laboratory for Novel Software Technology since 2010.

Chew Lim Tan is a Professor in the Department of Computer Science, School of Computing, National University of Singapore. He received his B.Sc. (Hons) degree in physics in 1971 from University of Singapore, his M.Sc. degree in radiation studies in 1973 from University of Surrey, UK, and his Ph.D. degree in computer science in 1986 from University of Virginia, U.S.A. His research interests include document image analysis, text and natural language processing. He has published more than 460 research publications in these areas. He is a fellow of the International Association of Pattern Recognition (IAPR). He is also a senior member of IEEE.