Local Resultant Gradient Vector Difference and Inpainting for 3D Text Detection in the Wild

Dajian Zhong[1], Palaiahnakote Shivakumara[2], Lokesh Nandanwar[2], Umapada Pal[3], Michael Blumenstein[4] and Yue Lu[1, 5]

[1]Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai, China.
[2]Department of Computer System and Technology, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia.
[3]Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India.
[3]University of Technology Sydney, NSW, Australia.
[5]School of Communication and Electronic Engineering, East China Normal University, Shanghai, China,
Email: djzhong@stu.ecnu.edu.cn, shiva@um.edu.my, lokeshnandanwar150@gmail.com, umapada@isical.ac.in, michael.blumenstein@uts.edu.au, ylu@cee.ecnu.edu.cn

Abstract. 3D text appearing in natural scene images is common due to 3D cameras and the capture of text from different angles, which presents new problems for text detection. This is because of the presence of depth information, shadows, and decorative characters in images. In this work, we consider those images where 3D text appears with depth, as well as shadow information for text detection. We propose a novel method based on Local Resultant Gradient Vector Difference (LRGVD), inpainting and a deep learning model for detecting 3D as well as 2D texts in natural scene images. The boundary of components that are invariant to the above challenges is detected by exploring LRGVD. The LRGVD uses gradient magnitude and direction in a novel way for detecting the boundary of the components. Further, we propose an inpainting method in a new way for restoring the character background information using boundaries. For a given region and the input image, the inpainting method divides the whole image into planes and then propagates the values in the planes into the missing region based on posterior probabilities and neighboring information. This results in text regions with false positives. Then the Differential Binarization Network (DB-Net) is proposed for detecting text irrespective of orientation, background, 3D or 2D, etc. Experiments conducted on our 3D text images and standard datasets of natural scene text images, namely, ICDAR 2019 MLT, ICDAR 2019 ArT, DAST1500, Total-Text, SCUT-CTW1500, show that the proposed method is effective in detecting 3D and 2D texts in the images.

Keywords: Local gradient vectors, Boundary points, Text detection, Inpainting, Deep learning, 3D text detection.

1 Introduction

Most day-to-day activities are recorded by both 2D and 3D cameras, especially for surveillance and monitoring applications. For example, controlling traffic jams, parking of vehicles, overcrowding, disaster management, tracking persons during marathons, sports, exhibitions and processions etc., [1, 2]. At the same time, 2D cameras are also used for the same purposes. In addition, when we capture 3D text, such as a building name, street names using a 2D camera at different angles, 3D text is displayed with depth and shadow information. This work considers those images where 3D text appears with depth and with shadow information for text detection. As a result, one can expect a mix of 2D and 3D text images. In this situation, we need a method that can understand the content irrespective of 2D and 3D text. This work considers 3D text in 2D video frames and natural scene images for text detection. The input for this work is 2D natural scene images containing 3D text. In the case 2D image containing 3D text, we can see shadow information, if the image is captured at different angles (oblique), while for 2D text, shadows are not visible. If the 3D text image is captured in an orthogonal direction (head-on), the 3D text appears as 2D text.

There are some methods [3, 4] developed in the past, but most of them are limited to 2D text detection, and not 3D text in natural scene images. Therefore, when the data includes 3D text images, the performance of the existing methods degrades. This is due to the presence of shadows, depth and decorative characters in the 3D text. These adverse factors pose different challenges such as loss of vital information, degradation, distortions created by oblique angles, the presence of shadows, ornamental characters, etc. Text is an important clue for understanding the content at the semantic level as it gives meaning, which is close to the content and hence, it fills the gap between high-level and low-level features [3, 4, 5, 6]. Therefore, achieving better results for text detection irrespective of 2D and 3D texts is an ongoing challenge for the document analysis community. Thus, it requires an intelligent and effective method that can cope with the challenges of 2D and 3D text detection in natural scene images.

It is evident from the illustration shown in Fig. 1(a)-(c) that the existing text detection method, CRAFT [3], which is a character region awareness-based text detection method, does not apply proper bounding boxes for the 3D text as shown in Fig. 1(a). Similarly, the DB-Net [4], which is a differential binarization-based text detection method misses text in the 3D image as shown in Fig. 1(b). On the other hand, the proposed method is capable of detecting 3D text as shown in Fig. 1(c). This is because the proposed method involves a new idea of detecting boundaries of the text component based on local resultant gradient vector difference, and restoring character background information through an inpainting approach, which is robust to 2D and 3D text. It can be seen in Fig. 1(c) that boundary detection results remove most of the non-text information, and inpainting restores the shape of text components. In this way, the proposed method makes a substantial difference compared to the state-of-the-art methods.

It is noted from the above illustration that the proposed method comprises the image processing topics of the gradient, which is explored to develop a new approach called the local resultant gradient vector difference for boundary point detection. Similarly,

the pattern recognition topics of clustering for separating boundary points, which represent text from non-text components, and the computer vision sub-area utilizing an interpolation-based inpainting approach is used for predicting missing character pixels. Furthermore, the machine learning theme of deep learning is applied for text detection, irrespective of text type and effects of depth and shadow information. Overall, the proposed work involves image processing, pattern recognition, computer vision and machine learning, which are part of the broader field of artificial intelligence to achieve the best results for 3D text detection in this work.

(a) Results of CRAFT [3] (b) Results of DB-Net [4] (c) Results of the proposed method

Fig. 1. Motivation for proposing 3D text detection in natural scene images

The key contributions of the proposed work are as follows. To the best of our knowledge, there is no work for detecting 3D and 2D text in the literature, and thus the specific contributions are as follows: (1) Exploring the concept of local resultant gradient vector difference for finding boundary points of text is new for text detection in natural scene images. (2) Proposing an inpainting approach for restoring background information of characters such that character shapes of text can be restored using boundary points is novel. (3) Exploring a differential binarization network for accurate text detection, regardless of 2D and 3D text.

The rest of this paper is organized as follows. In Section 2 we review the relevant literature, classifying them into three groups: Regression / anchor-based methods, segmentation-based methods, and hybrid methods. Section 3 describes the main methodological part of the proposed work. In Section 4, we present the quantitative and qualitative experimental results. Ablation studies and comparative results are also included in this section. Finally, Section 5 draws some conclusions and outlines perspectives for future work.

2 Related Work

The methods based on CNNs are popular for 2D text detection in natural scene images, and images extracted from videos as well. It is demonstrated that these methods are capable of addressing almost all the challenges of natural scene text detection. The methods can be categorized as regression/anchor-based, segmentation-based methods,

and hybrid methods. Similarly, a few methods are proposed for addressing the challenges of 3D text detection in video.

2.1 Regression / Anchor-Based Methods

These models are proposed with the inspiration of object detection by treating the whole text segment as an object for text detection in the images.

Liu et al. [7] proposed a method for Fast Oriented Text Detection (FOTS) in natural scene images. The method uses a unified model for the detection and recognition of text in natural scene images. However, the performance of the method depends on the way the model integrates the steps. Liu et al. [8] presented a robust curve text detection method in natural scene images based on conditional spatial expansion. Since the main focus of the method is 2D text detection, the method may not be effective for the images of both 2D and 3D texts. He et al. [9] proposed a method for multi-oriented and multilingual text detection in natural scene images based on direct regression. The method is sensitive to arbitrarily shaped text lines. Cheng et al. [10] introduced a direct regression scene text detection approach in natural scene images based on positive-sensitive segmentation. The regression concept used in this work may not be effective for the images of different types of texts, such as 2D and 3D text and color bleeding. Dai et al. [11] developed a model for arbitrarily shaped scene text detection based on progressive contour regression. The method proposes three stages to extract correct contours for improving text detection performance. The approach may not be robust because the success depends on the success of the three stages. Li et al. [12] proposed a method to predict rectangles and the text center region for bilingual text detection in natural scene images. The focus of the method is particular languages, and the scope does not include 3D text detection in natural scene images. Liu et al. [13] proposed a model for curved text detection in natural scene images based on regression analysis, which explores transverse and longitudinal sequence connections for improving text detection performance. The method may not be effective for degraded images.

The main weakness of regression-based methods is the use of a rigid reference (anchor) for text detection in natural scene images. Due to this constraint, regression-based methods report poor results for the images of irregular and arbitrarily oriented text. To overcome this problem, anchor-based methods have been introduced to improve text detection performance. Deng et al. [14] proposed a real-time scene text detector using a learned anchor. Sheng et al. [15] presented a single shot-oriented scene text detector with learnable anchors. Hou et al. [16] described a method for scene text detection using a hidden anchor mechanism. However, defining anchors for different situations is challenging.

In brief, when there is a problem in defining correct anchors for the text because of confusion between text and shadow pixels in 3D images, there is a high chance of misclassifying non-text as text for predicting subsequent anchors. Therefore, these methods may not be effective for 3D text detection.

2.2 Segmentation-Based Methods

Regression and anchor-based methods are not adequate and accurate for handling curved and short text in natural scene images. This has motivated the development of segmentation-based methods, which extract the information at the pixel and character levels. Since the methods focus on pixel and character levels, they can be robust to arbitrary orientation, short text, and irregular-sized text.

Baek et al. [3] presented a method called Character Region Awareness for Text Detection (CRAFT). The approach may not work well for images affected by blur and degradations because character segmentation is not easy in blurred images. Liao et al. [4] proposed a method for text detection in natural scene images based on the Differentiable Binarization (DB) concept. The method works well for high-quality scene images but not for complex backgrounds and low contrast images. Based on a Progressive Scale Expansion Network (PSENet), Wang et al. [17] used a method for text detection in natural scene images. The approach is not robust to noise and distorted images because the noise and blur affect progressive scale expansion. Tang et al. [18] presented an approach for detecting dense and arbitrarily oriented text lines in natural scene images. The method may not provide stable and reliable results for dense arbitrary-oriented scene text in the images. Liu et al. [19] proposed a mask tightness text detector for arbitrarily shaped scene text detection. The performance depends on choosing relevant masks according to different situations. Dai et al. [20] described a method for curved text detection in natural scene images based on multi-scale context-aware feature aggregation. Although the method considers multi-lingual text images for detection, it does not consider images of 3D text or text with shadow information. Zhang et al. [21] discussed a technique for arbitrarily oriented text detection in natural scene images by exploring the omni-directional pyramid mask proposal network. However, the method may not be effective for dense text and short text containing two-three characters. Xing et al. [22] introduced an approach based on a convolutional character network for text detection in natural scene images. The character-based model may not be robust for noisy and blurred images. Liu et al. [23] proposed a method based on context attention and a repulsive text border. The context information can be extracted when the text line contains several words. Zhang et al. [24] used a deep relational reasoning graph network for arbitrarily shaped text detection in natural scene images. The method is said to be computationally expensive due to use of reasoning graph network. Wang et al. [25] proposed an efficient and accurate arbitrarily oriented shape text detector with a Pixel Aggregation Network (PAN). However, this method may not work well for the images of dense text lines where there is non-uniform spacing between the text lines. Liao et al. [26] developed a model called Mask TextSpotter for scene text detection and recognition in natural scene images. The method follows an instance segmentation approach for detecting and recognizing characters. However, the method may be sensitive to low contrast and low-resolution images.

In summary, when the images contain text with shadows and are affected by perspective distortion due to oblique angles, extracting features that represent a character is challenging. In addition, sometimes when the shadow pixels share the properties of characters, segmentation-based methods fail to extract actual text information, and

hence the performance of the methods degrade.

2.3 Hybrid Methods

Although segmentation-based methods are robust to certain challenges, the use of deep learning models depends heavily on the number of samples and a large number of parameters. These hard constraints limit the generalizability of the methods. To find a solution to this problem, hybrid methods are proposed that integrate the merits of pixel/character level information by extracting handcrafted features and using deep learning models.

Wang et al. [27] used a hybrid method, which combines the advantage of segmentation and regression-based methods to overcome the limitations of the previous methods. However, the performance of the approaches depends on the way the method uses the advantages of each concept. Roy et al. [28] presented a text detection method from multi-views of the scene images based on the Delaunay Triangulation concept. Xue et al. [29] introduced a method for arbitrarily oriented text detection in low-light natural images. Nag et al. [30] proposed a unified method for detecting text in marathon runner and sports player images. Although the techniques explore the combination of feature extraction and deep learning models for achieving better results, sometimes, obtaining consistent and stable results for different datasets becomes difficult.

The review of the above methods shows that the existing models are capable of addressing the challenges of arbitrarily oriented text, different shaped text, and low-light text detection in natural scene images. However, none of them focuses on 3D text detection in images. In reality, the same natural scene images may have 3D text along with 2D text, especially when the names of buildings are captured from different angles. To address the 3D text detection problem, there are existing approaches that have been developed [31, 32]. For example, one method uses the combination of local gradient difference and cloud of line distribution for feature extraction and subsequently a neural network is introduced for 2D and 3D text image classification [31]. The approach aims to classify the 2D and 3D scene text images such that an appropriate method can be used for improving text detection performance. The method explored the combination of wavefront concepts and deep learning for text localization in 3D video [32]. The approach uses gradient vector flow concepts for dominant point detection and the wavefront concept for text candidate detection. Furthermore, the adaptive B-spline polygon curve-based network is applied for text detection in 3D video. This has motivated us to propose 3D and 2D text detection. Achieving stable and reliable results for 2D and 3D text is not addressed by the above-mentioned method. This is due to the loss of shape of the characters while detecting text candidate points using the wavefront concept.

Hence, in this work, we propose a new method for 3D and 2D text detection by exploring the local resultant gradient vector difference and an inpainting approach in a novel way. When the image contains both 2D and 3D text, the boundary of text is a common feature for both 2D and 3D texts. It is also true that the pixels of shadows have low values compared to boundary pixels of text. In addition, it is noted that the gradient and direction information of text pixels are similar in the images [31]. These observations motivated us to introduce Local Resultant Gradient Vector Difference (LRGVD)

[33] for finding the boundary points of text in the images. Due to the complexity of the problem, there is a chance of losing boundary points and character shapes. Inspired by the inpainting approach, which restores the missing information with the help of neighboring information and plane generation for the given region [34], we explore the same concept for restoring missing pixel information from text components using boundary points. This step results in a region of interest which contains text information. Therefore, this process reduces the background complexity of the problem because it eliminates most of the non-text information in the background text and shadow information. With this advantage and motivated by the ability of deep learning models for addressing challenges of text detection in natural scene images, we explore the deep Differential Binarization Network (DB-Net) [4] for text detection by feeding text regions as inputs in this work.

3 Model for 3D Text Detection

As discussed in the previous section, the proposed approach consists of three steps, namely, local resultant gradient vector difference for the detection of boundaries of the text components in the images, an inpainting method for restoring character background information, and deep learning for text detection as shown in Fig. 2. Since the boundary of the text is independent of 2D and 3D texts, and the boundary of the characters share almost the same properties, we propose to explore Local Resultant Gradient Vector Difference (LRGVD) for boundary detection. This is understandable because the pixels which represent characters almost have uniform values and hence one can expect the gradient and direction of the text boundary exhibiting a unique relationship [31]. In the case of text pixels, they exhibit high values, whereby for non-text pixels, they demonstrate low values.

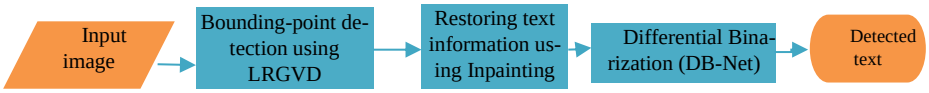


Fig. 2. Block diagram of our proposed model

It is evident from Fig. 3, where one can see a wide gap between the values of LRGVD of the boundary of text and non-text pixels. This is due to the neighboring gradient values of the boundary points of text pixels sharing almost the same high values while the neighboring gradient value of boundary points of non-text pixels do not. Therefore, the LRGVD outputs high values for the boundary points of text pixels and low values for the boundary points of non-text pixels. The motivation to propose LRGVD in contrast to normal gradient values is that LRGVD can withstand the adverse effects of depth and shadow information in 3D text images. This is because the LRGVD is estimated using gradient differences of neighboring pixels while the gradient is estimated using pixel differences of neighboring pixels. It is illustrated in Fig. 4, where it can be seen that LRGVD removes almost all non-text components while the conventional gradient does not. This demonstrates that the proposed LRGVD is robust to low contrast and complex backgrounds compared to the conventional gradient.

Based on the above-mentioned proposition that LRGVD gives high values for text boundary pixels and low values for non-text boundary pixels, the proposed method calculates the rate of change score. The proposed method employs k-means clustering for classifying high values as a text cluster, which results in boundaries of text components. However, due to adverse factors of 3D decorative characters and the presence of shadows, the boundary detection step may lose some of the pixels, which leads to a loss of character shape.

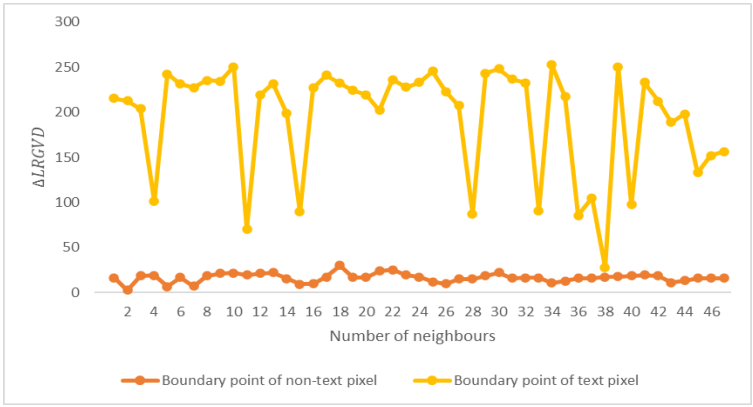


Fig. 3. The distribution of $\Delta LRGVD_{(p,k)}$ for the boundary of text and non-text pixels.

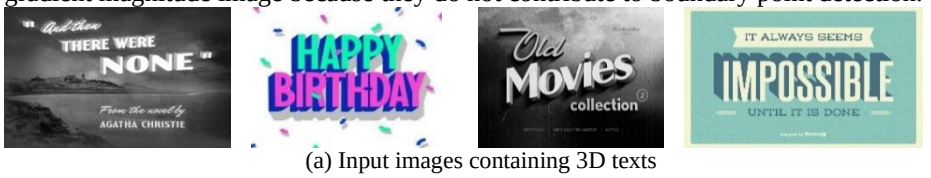
Input 3D text image Normal gradient LRGVD

Fig. 4. The effect of the proposed LRGVD over conventional gradient operations

To restore character background information and missing boundaries, we explore an inpainting approach because inpainting approaches use background information in the generated planes of the input image and neighboring information of boundaries to restore pixels. This step results in characters with a clear shape. Sometimes, due to the presence of background objects, the above process produces false text components. Therefore, to improve text detection performance by removing false positives, we are inspired by the impressive performance of the Differential Binarization Network (DB-Net) for text detection. In natural scene images, we explore the DB-Net which fixes the exact bounding box for the text line of any orientation by considering the results of inpainting as inputs.

3.1 Local Resultant Gradient Vector Difference (LRGVD) for Boundary Points Detection

As discussed in the previous section, boundary detection is important to reduce the background complexity of images for 3D text detection. For each input image shown in Fig. 5(a), the proposed method obtains the horizontal gradient, vertical gradient and gradient magnitude images, say, $grad_x$, $grad_y$, and $grad$ as shown in Fig. 5(b)-(d), where one can see prominent edge information is sharpened especially in 5(d), which uses both horizontal and vertical gradient information. If the gradient magnitude image contains any zero values, the proposed method eliminates all such zero values from the gradient magnitude image because they do not contribute to boundary point detection.



(a) Input images containing 3D texts

(b) Horizontal gradient image

(c) Vertical gradient image

(d) Gradient magnitude image

(e) Detected boundaries of character components

Fig. 5. Steps of the boundary detection process on text components

In order to arrange all non-zero gradient values, we construct a KD tree such that the non-zero gradient values are indexed in order, which facilitates the determination of nearest neighbors for boundary point detection. For each pixel *p*, in the gradient magnitude image $grad$, the proposed method finds the resultant gradient using the gradient vector of the *p*, and the gradient vectors of its neighbors chosen from horizontal and

vertical gradient images as defined in Equation (1). The number of nearest neighbors is defined as k = 1, 2, 3, ……to 50 (maximum).

$$\overrightarrow{LRGVD}_{(p,k)} = \sum_{q=1}^{k} \overrightarrow{grad}_{p,q} \tag{1}$$

where $\overrightarrow{grad}_{p,q}$ represents the gradient between the two points. The distance and angle between the two points of $\overrightarrow{grad}_{p,q}$ are defined by Equation (2) and Equation (3).

$$\left|\overrightarrow{grad}_{p,q}\right| = \left|grad_{x,q} - grad_{x,p}\right| + \left|grad_{y,q} - grad_{y,p}\right| \tag{2}$$

$$\theta = tan^{-1}\left(\frac{C_{q,y} - C_{p,y}}{C_{q,x} - C_{q,x}}\right) \tag{3}$$

where $grad_x$ and $grad_y$ are defined in Equation (4) and Equation (5). $(C_{p,x}, C_{p,y})$, and $(C_{q,x}, C_{q,y})$ represent the spatial coordinates of point p and point q.

$$grad_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} * A \tag{4}$$

$$grad_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} * A \tag{5}$$

where '*' means convolution operation on the input image A.

For each k, the proposed method finds the resultant gradients with the same pixel, *p*. This results in a Local Resultant Gradient Vector Difference (LRGVD) matrix, which contains resultant gradient vectors of each pixel in the image. Then the rate of change in LRGVD for each pixel, *p*, is calculated as defined in Equation (6), which is the absolute difference between the Local Resultant Gradient Vector of pixel *p* and its different *k* neighbors, denoted as $\Delta LRGVD_{(p,k)}$. When we observe the values in $\Delta LRGVD_{(p,k)}$, it is found that the values can be divided into two sub-groups because there is a big difference between high and low values which can be verified from the line graphs plotted for the number of neighbors *vs* $\Delta LRGVD_{(p,k)}$ values. Since the proposed work requires the boundary of text and there is a wide gap between the boundary of text and non-text, we propose to use k-means clustering with k=2. The cluster that gives the highest mean is considered as the boundary of the text cluster, which we call the boundary of text in this context. The reason to use k-means clustering for obtaining boundary points of text clusters is the following. Since it is true that the pixels represented by low gradient difference values do not contribute much to text detection, it is necessary to eliminate low gradient difference values. This observation motivated us to use K-means clustering with K=2 for classifying boundary pixels represented by high gradient difference values, which output two clusters. The effect of the above process can be seen in Fig. 5(e), where we can see almost all boundaries of the text are detected for different cases. Therefore, the boundary detection step reduces the background complexity of the problem by removing almost all non-text components. This is understandable because the step considers taking advantage of the gradient vector of the pixels and calculating the resultant gradient on each pixel with respect to its k-neighbors for detecting boundaries. For complex scenarios, gradient magnitude alone or gradient vector alone is not sufficient to handle this situation.

$$\Delta LRGVD_{(p,k)} = \left| \left| \overrightarrow{LRGVD}_{(p,k)} \right| - \left| \overrightarrow{LRGVD}_{(p,k+1)} \right| \right|, k = 1, 2, \ldots, K-1 \quad (6)$$

where $\left| \overrightarrow{LRGVD}_{(p,k)} \right|$ represents the resultant magnitude of $\overrightarrow{LRGVD}_{(p,k)}$.

3.2 Inpainting Method for Restoring Text Information

It is noted from the results of boundary detection shown in Fig. 5(e) that the information of characters, as well as the boundaries of a few characters, are missing. This is due to the complex background and low contrast information of the text. However, boundary points detection reduces the complexity of the image and the effect of depth and shadow as shown in Fig. 5(e), where the step outputs individual characters with clear boundary points. It is observed that the pixels of the characters share almost the same values in the text. In addition, the background of the character usually has homogeneous regions compared to that of non-text components. These two observations are common for any type of text. These insights motivated us to introduce the inpainting approach for restoring background information of characters such that the character shape can be preserved. This is because the inpainting approach uses neighboring information for restoring missing pixels through an interpolation technique, and hence the inpainting methods works well for the image containing homogeneous regions.

To restore such information, inspired by the method in [34], where the background information is filled in place of the obstacle of the natural scene image, we explore the same method for a different situation, where the background information of the character component is to be restored.

(a) Connected component bounding boxes on boundary images

(b) Filtered bounding boxes by expanding and merging the overlapping boxes

(c) The results of the inpainting approach.

Fig. 6. Restoring text information using the inpainting method.

Since the boundary point detection step preserves the structure of characters as shown in Fig. 5(e), we proposed to use a connected component approach for extracting the character components. The connected component approach traverses the boundary

points of the characters and finds maxima and minima in the X and Y directions. This results in the locations of four corners of each character component. The locations of the four corners are used for fixing the bounding box and the extraction of character components. The advantage of this approach is that in the case of the boundary of characters missing pixels (if there are any disconnections), the rectangular bounding boxes cover the full character information. In this way, the connected approach helps us to extract character components irrespective of size, text and non-text as shown in Fig. 6(a), where we can see bounding boxes for the components.

Fig. 7. The effect of the proposed inpainting approach for restoring the shape of characters using boundary points

In order to employ the inpainting approach, the proposed method merges the small components as one single component based on overlapping information of bounding boxes as shown in Fig. 6(b). The merged components are required to restore the missing background information, and along with the input images, they are fed to the inpainting approach. The inpainting approach divides the input image into a number of planes and then the regular structure of planes is propagated into the region of merged connected components based on a posterior probability. The posterior probability of the missing background pixels is calculated using the information of planes and neighboring information of pixels of the merged connected components. The effectiveness of the inpainting approach is shown in Fig. 6(c), where it is seen that the background information of the characters is restored. The advantage of this step is that it results in a text region which can be considered as a region of interest. The same is fed to a deep learning model for 3D and 2D text detection in the images. Since the background information of the input image is removed, the convolutedness of the problem is reduced. As a re-

sult, it helps in addressing the problems of 3D text detection. The effect of the inpainting approach can be seen in Fig. 7, where one can see the background information is restored using the results of boundary point detection for the 3D text images.

3.3 Deep Learning for Text Detection in 3D Images

Motivated by the impressive performance of the deep learning model (which is a Differential Binarization Network (DB-Net) [4]) for text detection in 2D natural scene images, we use the same for 3D text detection. In this work, we consider the region of interest, which is given by the inpainting method as the input. To overcome the limitation of DB-Net on 3D text detection, the region of interest is input rather than the full image for text detection. In this work, we perform fine-tuning on the pre-trained model weights, where ResNet18 is used as the backbone network and trained on the ICDAR 2019-MLT dataset [35]. The DB-Net is a Differential Binarization net, which performs binarization with an approximate step function as defined in Equation (7).

$$f_{x,y} = \frac{1}{1 + e^{-\alpha(\rho_{x,y} - \kappa_{x,y})}} \tag{7}$$

where $f_{x,y}$, $\rho_{x,y}$ and $\kappa_{x,y}$ denote the binary segmentation prediction, probability mask and border mask, respectively. Here, α denotes the amplifying factor and is set to 50. This solves the problem of non-differentiability of the standard binarization methods. In this method, the input image is passed into a feature-pyramid encoder, up-sampled to equal scales and concatenated to produce a final feature, which is used to perform prediction of the 2 types of masks, namely the probability mask $\rho_{x,y}$ and the border mask $\kappa_{x,y}$. Training supervision is applied on all the three masks, that is, $f_{x,y}$, $\rho_{x,y}$ and $\kappa_{x,y}$, and inference can be performed by directly fixing the bounding box around $f_{x,y}$. The effect of the text detection results for different cases can be seen in Fig. 6, where the proposed combination of regions of interest detection as well as the accurate detection of 3D text via the DB-Net are illustrated. This is the advantage of the proposed method.

4 Experimental Results

Since a standard dataset for 3D text detection is not available publicly, we created our own dataset from different sources, such as YouTube, movie posters, and the internet, where 3D text has depth and shadow information. The dataset created includes diversified images chosen from different 3D text images, and it contains a total of 400 images for experimentation. For creating the ground truth for our dataset, we follow the instructions and the steps used in [13] for fixing bounding boxes manually. For horizontal and simple oriented text lines, four clicks are sufficient to annotate the text, while for curved text, we need to use many clicks for polygonal bounding box. In order to increase the number of samples for training, we use augmentation techniques, namely, rotation, flipping, scaling etc. In order to demonstrate that the proposed method is effective for 2D text detection in natural scene images, we also considered benchmark datasets for experimentation on natural scene text, as follows.

ICDAR 2019-MLT [35]: It is constructed for evaluating text detection performance on multi-lingual text, namely, Arabic, Bangla, Chinese, Devanagari, English, French, German, Italian, Japanese and Korean. It provides 1000 images per language for training and 1000 images per language for testing. **SCUT-CTW1500 [3]**: This dataset provides arbitrary-shaped text-line natural scene images in English and Chinese scripts. For experimentation, there are 1000 images for training and 500 images for testing. **Total-Text [3]**: This dataset provides images containing curved text similar to the CTW1500 dataset. However, most of the images contain English text lines. In this dataset, 1255 images are considered for training and 300 images for testing to perform experimentation. **ICDAR 2019 ArT [36]**: This dataset is a combination of the images of Total-Text, the CTW1500 datasets and Baidu Curved Scene Text, which was created for detecting arbitrary-shaped text in natural scene images. In total, the dataset contains 10,166 images, which is split into a training set of 5603 images and a testing set of 4563. **DAST1500 [18]**: This dataset considers dense and arbitrary-shaped text images, which contains 1538 images. Out of the 1538 images, 1038 were used for training, and 500 images for testing.

The reason to choose the above five standard datasets for experimentation is the following. The objective of the proposed work is not only to address the challenges of 3D text detection but also to address the other challenges like arbitrary orientation (CTW1500, Total-Text and ICDAR 2019 ArT datasets contain arbitrary orientations text), multiple scripts (ICDAR 2019 MLT is a multi-script-data), and dense text (DAST1500 contains dense text). In addition, these five datasets are complex and widely used for evaluating the performance of text detection compared to other datasets. Furthermore, it is observed that these datasets contain a few images comprising 3D texts. In the case of the CTW1500 dataset, it does not provide the ground truth for the words, rather it gives the ground truth for the whole text line. Therefore, for experimentation, the proposed method uses an additional step for merging the bounding boxes of the words to fix a single bounding box for the whole text line.

For measuring the performance of the proposed method, we use standard measures and an evaluation scheme, namely, Recall (R), Precision (P) and F-measure (F) for all the experiments in this work. The threshold for intersection-over-union (IoU) for classifying true and false positives is 0.5 according to the standard evaluation scheme [35, 36]. In the same way, to show that the proposed method is robust, we use the following state-of-the-art natural scene text detection methods for a comparative study. The CRAFT [3] approach, which is Character Region Awareness for Text Detection (CRAFT), the DB-Net [4], which is a Differential Binarization Network, the PSENet [17], which is a Progressive Scalable Expansion Network, the CTD-Net [13], which is a Curved scene Text Detection network, and the Mask-TextSpotter [26], which is an end-to-end text detection and recognition method, are used for a comparative study with the proposed method. The motivation to consider the aforementioned methods for a comparative study is that these approaches are state-of-the-art methods and address several challenges, which are the same for 3D text detection.

4.1 Ablation Study

In the proposed method, boundary detection and the inpainting steps are the key ones for achieving better results for 3D text detection. To show that the key steps are effective and contribute to achieving the results, we conducted the following experiments.

Experiment-(i): The output of the boundary of components is supplied to the deep learning step for text detection without an inpainting step, and the results of recall, precision and F-measure on our dataset are reported in Table 1. It is noted from the results of experiment (i), especially for F-measures, that the proposed method with a boundary reports reasonable results, but it is not the best compared to the proposed method (baseline). Experiment (ii): In the same way, we conduct experiments without boundary detection and inpainting by passing the input images directly to the deep learning approach for text detection. The results are reported in Table 1, where one can see the results are not higher than the proposed method. This shows that the deep learning step alone may not be sufficient to address the challenges of 3D text detection in images. The reason for the poor results of experiment (ii) is that the performance of deep learning depends on the number of samples and tuning parameters and hence it lacks the ability of generalization in contrast to the proposed combination of features and deep learning. However, deep learning has the ability to achieve consistent results for different problem complexities. Therefore, the proposed approach considers the advantages of both feature extraction (boundary point detection as well as inpainting) and deep learning for achieving impressive results for 3D text detection in this work, as reported in Experiment (iii). In summary, we can infer that the key steps contribute equally, and the key steps alone are not adequate to achieve the best results as the proposed method.

Table 1. Analyzing the effectiveness of different experiments proposed in this work for text detection in 3D video images

Experiments	(i)	(ii)	(iii) Proposed method
Precision	73.0	72.1	79.4
Recall	13.6	73.2	73.1
F-measure	23.0	72.6	76.1

4.2 Evaluating the Proposed Text Detection approach

Qualitative results of the proposed method for 3D text detection on sample images from our dataset are shown in Fig. 8. One can see that the proposed approach detects text well for all the images. This shows that the proposed method is capable of handling 3D text in the images. The same inferences can be drawn from the quantitative results of the proposed method and existing methods reported in Table 2. It is observed from Table 2 that the proposed method is better than the existing methods in terms of precision and F-measure. This indicates that the proposed method is reliable compared to the existing methods for 3D text detection in the images. However, sometimes, for very low contrast images, the LRGVD misses some text pixels. In this case, if the inpainting step does not restore it, the proposed method misses text in the images and hence the recall is low compared to precision and F-measure. The main reason for the lower results of the existing methods is that the methods are developed for 2D text detection

but not 3D text detection. In addition, the existing methods lack a generalization ability compared to the proposed method because the success of existing methods depends on the number of samples and too many tuning parameters. On the other hand, since the proposed method involves boundaries of text detection and inpainting, which are invariant to the effects of 3D text, the performance of the proposed method is superior.

Fig. 8. Text detection results of the proposed method on our 3D text dataset.

Table 2. Comparative performance with the existing methods on our dataset

Methods	Precision	Recall	F-Measure
PSENet [17]	72.6	77.7	75.1
DB-Net [4]	73.4	77.2	75.3
CRAFT[3]	70.6	**79.0**	74.6
CTD [13]	70.3	77.1	73.5
MaskTextSpotter [26]	71.4	77.7	74.4
Proposed Method	**79.4**	73.1	**76.1**

To show the proposed method is effective for 2D text detection, we conducted experiments on five standard datasets of natural scene text detection. Qualitative results of the proposed method for the images of the respective five standard datasets are shown in Fig. 9, where it is noted that the proposed method detects text of different orientations and complexities accurately. Therefore, one can assert that the proposed approach is independent of text type and image type. Quantitative performance results of the proposed and existing methods on different standard datasets are reported in Table 3, where it is observed that the proposed method is better than existing methods for all the datasets in terms of precision and F-measure except for the Total-Text and ICDAR 2019 MLT datasets. Sometimes, the step of boundary detection and inpainting of the proposed method misses text pixels in the images, which do not have sufficient contrast, and hence the recall is lower than the precision and F-measure. Since DB-Net involves a powerful differential binarization network, which is capable of extracting contextual information from the text, it obtains the best recall for a few of the datasets. However, the precision is lower than for the proposed method. This shows that the proposed method is effective and reliable for 2D text detection in natural scene images.

The reason for the lower results of the existing methods includes the inherent limitations such as the confined scope of the methods, objective of the methods, constraints on training, the number of samples, etc. On the other hand, although the proposed method is developed for 3D text detection, it outperforms the existing methods. This is because of key steps which are invariant to text type and the effect of depth and shadow information in the images.

Fig. 9. Qualitative results of the proposed method for the different standard datasets of natural scene text detection.

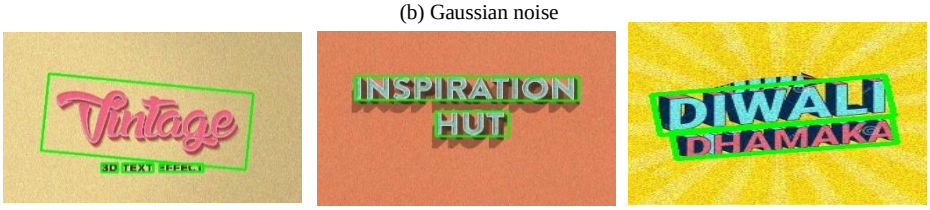
Table 3. Comparative performance of the proposed and existing methods on different benchmark datasets of 2D text detection in natural scene images

Methods	SCUT-CTW1500	Total-Text	ICDAR2019Art	DAST1500	ICDARMLT2019

	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
PSENet [17]	79.7	84.8	82.2	84.0	77.9	80.8	81.1	57.5	67.3	74.6	50.1	60.0	73.5	59.5	65.8
DB-Net [4]	80.2	**86.9**	83.4	87.1	**82.5**	**84.7**	56.0	**69.9**	62.2	71.4	65.2	68.2	84.5	**70.0**	**76.6**
CRAFT[3]	86.0	81.1	83.5	87.6	79.9	83.6	79.4	66.6	72.4	61.9	**88.2**	72.8	81.4	62.7	70.8
CTD[13]	77.4	69.8	73.4	74.0	71.0	73.0	-	-	-	84.5	77.1	**80.6**	-	-	-
MaskText Spotter[26]	80.7	84.3	82.5	81.8	75.4	78.58	-	-	-	80.1	69.5	74.4	-	-	-
Proposed Method	**86.3**	80.5	**83.8**	**88.2**	76.1	81.7	**81.6**	67.9	**74.1**	**93.1**	64.7	76.3	**90.2**	65.5	75.9

(a) Input 3D text images

(b) Gaussian noise



(c) Speckle noise

(d) Poisson noise

Fig. 10. The proposed text detection method on different noisy images

To show the proposed method is robust to noise, the method is tested on noisy images created by adding Gaussian, Speckle and Poisson noise at different levels (mean and variance are set to 0 and 0.01, respectively for applying the noise) as shown in Fig.10(a)-(d). It is observed from Fig. 10 that the proposed method can cope with the adverse effects of different noise and distortions to some extent.

When the boundary point detection step fails to preserve the shape of the characters due to complex backgrounds, low contrast and color bleeding, there is a chance of losing text as shown in Fig. 11(a)-(b). However, the situation mentioned above occurs seldomly, and hence it does not much affect the overall performance of text detection in the proposed method. This is the limitation of the proposed method, and it is beyond

the scope of the proposed work. In the same way, it is noted from the results shown in Fig. 11(c) that for the images where text is embedded over another text component, and text is affected by blur and color bleeding due to 3D effects, the proposed method does not detect the text accurately. The key reason for poor performance of the proposed method is that the boundary point detection and inpainting approach are sensitive to color differences and blurry information. Therefore, there is scope for expanding the proposed work for making it robust to different adverse situations. One way to overcome this challenge is to combine text detection with natural language processing concepts to differentiate text based on semantics. This will be our future work.

(a) Input images (b) The results of the inpainting approach

(c) Limitations of the proposed method for 3D text detection

Fig. 11. Erroneous results (missing text instances) of the proposed method for 3D text detection

5 Conclusions and Future Work

In this work, we have proposed a new method for 3D and 2D text detection in natural scene images. The 3D text images are extracted from 3D movies and cartoons. The proposed method consists of two new steps, namely, boundary detection by the Local Resultant Gradient Vector Difference (LRGVD) and an inpainting method for restoring character background information using boundaries of text components. The results of the inpainting method, which forms the region of interest, are passed to the deep learning model for text detection. The LRGVD explores gradient magnitude and gradient direction in a novel way for detecting successful boundaries of the text components. In the case of the inpainting approach, the method obtains planes from the input images, the background information of planes and neighboring information of the boundary points for restoring character background information. For text detection, the proposed method uses DB-Net by feeding the results of the inpainting approach for text detection. Experimental results on our 3D text dataset demonstrate that the proposed method outperforms the existing methods. The results on benchmark datasets of 2D text demonstrate that the proposed method is effective for 2D text detection in natural scene images. However, for the images affected by very low contrast, uniform color for both the

background and the text, the performance of the proposed method degrades, which will be the topic of our future research work.

Acknowledgments

This work receives partial support from the FRGS grant (FP104-2020), Ministry of Higher Education, Malaysia. The authors thank Swati Kanchan for assisting us to complete the experiments.

References

1. J. Xu, P. Shivakumara, T. Lu, C. L. Tan and S. Uchida, "A new method for multi-oriented graphics-scene-3D text classification in video", Pattern Recognition, pp 19-42, 2016.
2. W. Wei, J. Wu and C. Zhu, "Special issue on role of computer vision in smart cities", Image and Vision Computing, 2021.
3. Y. Baek, B. Lee, D. Han, S. Yun and H. Lee, "Character region awareness for text detection", In Proc. CVPR pp 9365-9374, 2019.
4. M. Liao, Z. Wan, C. Yao, K. Chen and X. Bai, "Real-time scene text detection with differentiable binarization", In Proc. Proc. AAAI, 2020.
5. X. Zhao, Z. Zhou, L. Li, L. Pei and Z. Ye, "Scene text detection based on fusion network", Int. J. Pattern Recognit. Artif. Intell. Vol. 35, No. 10, 2153005, 2021.
6. Z. Huang, P. Shivakumara, T. Lu, U. Pal, M. Blumenstein, B. Chetty and G. H. Kumar, "Improving ring radius transform-based reconstruction for video character recognition", Int. J. Pattern Recognit. Artif. Intell. 35(7): 2150023:1-2150023:24, 2021.
7. X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao and J. Yan, "FOTS: Fast Oriented Text Spotting with a Unified Network", In Proc. CVPR, pp 5676-5685, 2018.
8. Z. Liu, G. Lin, S. Yang, F. Liu, W. Lin and W. L. Goh, "Towards robust curve text detection with conditional spatial expansion", In Proc. CVPR, pp 7261-7270, 2019.
9. W. He, Z. Y. Zhang, F. Yin and C. L. Liu, "Multi-oriented and multi-lingual scene text detection with direct regression", IEEE Trans. IP, pp 5406-5419, 2018.
10. P. Cheng, Y. Cai and W. Wang, "A direct regression scene text detector with positive-sensitive segmentation", IEEE Trans. CSVT, pp 4171-4181, 2020.
11. P. Dai, S. Zhang, H. Zhang and X. Cao, "Progressive contour regression for arbitrary-shape scene text detection", In Proc. CVPR, pp 7389-7398, 2021.
12. J. Li, Y. Hao, W. Wang, T. Wang and Q. Li, "Scene text detection based on expanding the text center region for bilingual Tibetan Chinese", Int. J. Pattern Recognit. Artif. Intell., Vol. 35, No. 13, 2153007. 2021.
13. Y. Liu, L. Jin, S. Zhang, C. Luo and S. Zhang, "Curved scene text detection via transverse and longitudinal sequence connection", Pattern Recognition, pp: 337-345, 2019.
14. L. Deng, Y. Gong, X. Lu, Y. Lin, Z. Ma and M. Xie, "STELA: A real time scene text detector with learned anchor", IEEE Access, pp 153400-153407, 2019.
15. F. Sheng, Z. Chen, T. Mei and B. Xu, "A single shot-oriented scene text detector with learnable anchors", In Proc. ICME, pp 1516-1521, 2019.
16. J. B. Hou, X. Zhu, C. Liu, S. Sheng and L. H. Wu, "HAM: Hidden anchor mechanism for scene text detection", IEEE Trans. IP, pp 7904-7916, 2020.
17. W. Wang, E. Xie, X. Li, W. Hou, T. Lu, G. Yu and S. Shao, "Shape Robust Text Detection With Progressive Scale Expansion Network", In Proc. CVPR, pp 9328-9337, 2019.

18. J. Tang, Z. Yang, Y. Wang, Q. Zheng, Y. Xu and X. Bai, “SegLink++: Detecting dense and arbitrary-shaped scene text by instance-aware component grouping”, Pattern Recognition, 96, 2019.
19. Y. Liu, L. Jin and C. Fang, “Arbitrarily shaped scene text detection with a mask tightness text detector”, IEEE Trans. IP, 29, pp 2918-2930, 2020.
20. P. Dai, H. Zhang and X. Cao, “Deep multi-scale context aware feature aggregation for curved scene text detection”, IEEE Trans. MM, 22, pp 1969-1984, 2020.
21. S. Zhang, Y. Liu, L. Jin, Z. Wei and C. Shen, “OPMP: An omni-directional pyramid mask proposal network for arbitray-shape scene text detection”, IEEE Trans. MM, 2020.
22. L. Xing, Z. Tian, W. Huang and M. R. Scott, “Convolutional character networks”, In Proc. ICCV, pp 9126-9136, 2019.
23. X. Liu, G. Zhou, R. Zhang and X. Wei, “An accurate segmentation-based scene text detector with context attention and repulsive text border”, In Proc. CVPRW, pp 2344-2352, 2020.
24. S. X. Zhang, X. Zhu, J. B. Hou and C. Liu, “Deep relational reasoning graphs network for arbitrary shape text detection”, In Proc. CVPR, pp 9696-9705, 2020.
25. W. Wang, E. Xie, X. Song, Y. Zang, T. Lu, G. Yu and C. Shen, "Efficient and Accurate Arbitrary-Shaped Text Detection With Pixel Aggregation Network”, In Proc. ICCV, pp. 8439-8448, 2019.
26. M. Liao, P. Lyu, M. He, C. Yao, W. Wu and X. Bai, "Mask TextSpotter: An End-to-End Trainable Neural Network for Spotting Text with Arbitrary Shapes”, IEEE Trans. PAMI, pp 532-548, 2021.
27. Y. Wang, H. Xie, Z. Zha, M. Xing, Z. Fu and Y. Zhang, “ContourNet: Taking a further step toward accurate arbitrary shaped scene text detection”, In Proc. CVPR, pp 11750-11759, 2020.
28. S. Roy, P. Shivakumara, U. Pal, T. Lu, G. H. Kumar, “Delaunay triangulation-based text detection from multi-view images of natural scene”, Pattern Recognition Letters, pp 92-100, 2020.
29. M. Xue, P. Shivakumara, C. Zhang, Y. Xiao, T. Lu, U. Pal and D. Lopresti, “Arbitrarily-oriented text detection in low light natural scene images”, IEEE Trans MM, 2020.
30. S. Nag, P. Shivakumara, U. Pal, T. Lu and M. Blumenstein, “A New unified method for detecting text from marathon runner and sports player in video”, Pattern Recognition, 2020.
31. L. Nandanwar, P. Shivakumara, R. Raghavendra, T. Lu, U. Pal, D. Lopresti and N. B, Anur, “Local gradient difference features for classification of 2D-3D natural scene text images”, In Proc. ICPR, 1112-1119, 2021.
32. L. Nandanwar, P. Shivakumara, R. Ramachandra, T. Lu, U. Pal, A. Antonacopoulos and Y, Lu, “A new deep wavefront-based model for text localization in 3D video”, IEEE Trans. CSVT, 2021 (doi: 10.1109/TCSVT.2021.3110990).
33. J. Xie, Z. Xiong, Q. Dai, X. Wang and Y. Zhang, “A local gravitation-based method for the detection of outliers and boundary points”, Knowledge Based Systems, Vol 192, 2020.
34. J. B. Huang, S. B. Kang, N. Ahuja and J. Kopf, “Image completion using planar structure guidance”, ACM Transactions on Graphics, Vol. 33, pp 1-10, 2014.
35. N. Nayef, Y. Patel, M. Busta, P. N. Chowdhury, D. Karatzas, W. Khlif, J. Matas, U. Pal, J. C. Burie, C. L. Liu, and J. M. Ogier, “ICDAR 2019 robust reading challenge on multi-lingual scene text detection and recognition—RRC-MLT-2019”, In Proc. ICDAR, pp 1582-1587, 2019.
36. C. K. Chng, Y. Liu, Y. Sun, C. C. Ng, C. Luo, Z. Ni, C. M. Fang, S. Zhang, J. Han, E. Ding, J. Lu, D. Kartzas, C. S. Chan and L. Jin, “ICDAR 2019 Robust Reading Challenge on Arbitrary-Shaped Text-RRC-ArT”, in Proc. ICDAR 2019.