A New Deep Wavefront based Model for Text Localization in 3D Video

Lokesh Nandanwar, Palaiahnakote Shivakumara, Raghavendra Ramachandra, Tong Lu, Umapada Pal, Apostolos Antonacopoulos and Yue Lu

***Abstract*—With the evolution of electronic devices, such as 3D cameras, addressing the challenges of text localization in 3D video (e.g., for indexing) is increasingly drawing the attention of the multimedia and video processing community. Existing methods focus on 2D video and their performance in the presence of the challenges in 3D video, such as shadow areas associated with text and irregularly sized and shaped text, degrades. This paper proposes the first approach that successfully addresses the challenges of 3D video in addition to those of 2D. It employs a number of innovations, among which, the first is the Generalized Gradient Vector Flow (GGVF) for dominant points detection. The second is the Wavefront concept for text candidate point detection from those dominant points. In addition, an Adaptive B-Spline Polygon Curve Network (ABS-Net) is proposed for accurate text localization in 3D videos by constructing tight fitting bounding polygons using text candidate points. Extensive experiments on custom (3D video) and standard datasets (2D video and scene text) show that the proposed method is practical and useful, and overall outperforms existing state-of-the-art methods.**

***Index Terms*—Gradient vector flow, Wavefront, Deep learning, B-Spline curve fitting, Natural scene text detection, Text localization in 3D video.**

I. INTRODUCTION

Text localization in 3D video is an important topic for content-based video retrieval, particularly for annotating video based on semantics [1, 2]. It has attracted considerable research attention due to explosive growth of multimedia content which includes 2D and 3D video data, available [1, 2]. As a result, there is an increasing number of large repositories containing 2D, 3D video/images and multimedia content [1]. To ensure the robustness and accuracy of retrieval systems, text localization is vital as it provides significant semantic information for annotating video [3, 4]. Existing models focus on text localization in 2D video but not 3D video [3, 4]. Therefore, there is a need for a model that can work for both 2D and 3D video. Example of retrieval cases can include events extraction from 3D sports video, choosing a particular scene in a 3D movie, tracking and watching person behavior and interaction captured by 3D camera during exhibitions, processions, celebrations, etc. These situations motivated the authors to introduce the problem of text localization in 3D video in this work.

Since text localization in 2D images is a well-known problem, several methods can be found in the literature for addressing the challenges of 2D text localization. However, this is not the case for 3D text localization [5, 6]. In the work described in this paper, if an image contains text with shadows and depth, it is considered a 3D text image. Otherwise, it is considered a 2D text image. According to this characterization, 3D video of sports and movies was used to collect samples to be used in the experiments carried out in the context of the proposed work.

Due to the prevalence of 3D cameras, 3D movies and 3D sports broadcasts, the presence of 3D and 2D text in a single frame/image is becoming increasingly common. When the input to a system is a mix of 2D and 3D text images, existing models may not work well and hence performance degrades. This is arguably inevitable because the depth information in 3D text introduces shadows, and at the same time there seems to be an increase of use of decorative characters. The presence of shadow and decorative characters affects the shape of characters and causes non-uniform spacing between characters, words, and text lines. Hence the authors' effort to develop a model that can deal both with 2D and 3D text images, which can then be used to annotate video in order to retrieve specific events, and for understanding video and image irrespective of 2D and 3D text type.

It is noted that the uniform color of each character is one of the key properties for 2D text localization methods to differentiate between text and non-text pixels. However, for the images shown in Fig. 1, where shadows are present, this property does not work because shadow pixels have also almost uniform color. This is one of the main causes of the poor performance of existing 2D text localization methods. Similarly, due to the presence of decorative characters in 3D images as shown in Fig. 1, one can expect irregularly shaped text and non-uniform spacing between characters. Most of the 2D text localization methods use polygonal curve fitting for handling arbitrary orientated text, which works well only for uniformly sized and spaced text. Therefore, the performance of 2D text localization methods degrades in 3D images, where decorative characters exist [7].

Fig. 1 shows an example of 3D text (left hand side column) and one where both 3D and 2D texts are present. It can be seen in the results illustrated in Fig. 1 that the most prominent existing methods, such as the Character Region Awareness for Text Detection (CRAFT) [8], the Differential Binarization Network (DB-Net) [9], the Progressive Scale Expansion Network (PSENet) [10] do not detect text accurately within 3D video images. It should be noted that the above existing

- Lokesh Nandanwar and Palaiahnakote Shivakumara are with the Department of Computer System and Information Technology, University of Malaya, Kuala Lumpur, Malaysia. Email: lokeshnandanwar150@gmail.com, shiva@um.edu.my.
- Raghavendra Ramachandra is with Norwegian University of Science and Technology, Norway, Email: raghavendra.ramachandra@ntnu.no
- Tong Lu, Nanjing University, Nanjing, China, Email: lutong@nju.edu.cn
- Umapada Pal is with Indian Statistical Institute, Kolkata, India. Email: umapada@isical.ac.in.
- Apostolos Antonacopoulos is with the University of Salford, Salford, UK. Email: a.antonacopoulos@primaresearch.org
- Yue LU is with East China Normal University, Shanghai, China, Email: ylu@cs.ecnu.edu.cn

methods have been selected by the authors based on their ability to address different challenges like low contrast, low resolution, complex background, arbitrary orientation, irregular-sized text, etc. which are particularly relevant to the challenges of text localization in 3D video. These methods, despite having been developed using powerful deep learning models, miss part of the 3D text and for the part of the 3D text that has been identified, the bounding boxes are not correct as shown in the example on the left-hand side column in Fig. 1. Furthermore, even though the existing methods work quite well for 2D text images, they are not reliable for the images containing both 2D and 3D text (as can be seen in the right-hand side column example in Fig. 1).

The above shows that the existing methods are limited to 2D text localization. On the other hand, contrary, the method proposed in this paper localizes text successfully irrespective of whether it is 2D or 3D or a combination therefore in the same image as shown in Fig. 1. This is one of the significant differences between the proposed approach and the existing methods.

Figure 1. Examples of text localization in 3D (left) and 3D-2D combined (right) video images using the proposed and the other existing methods.

In summary, the contributions of the work described in this paper are as follows: (1) A new model is proposed, called Generalized Gradient Vector Flow (GGVF), for detecting dominant points in video by defining opposite arrow symmetry for the text pixels irrespective of challenges posed by shadow and decorative characters. (2) A new concept is proposed, the Wavefront, to filter out false dominant points, which result in potentially necessary candidate points and text patches. This is because the wavefront considers both the direction and its speed, enhancing the GGVF where only the direction is considered. (3) To handle the irregularly sized, arbitrary oriented text resulting from the presence of decorative characters, the proposed method introduces a new step called Adaptive B-Spline Polygon Curve fitting (ABS-Net) for accurately creating bounding boxes to describe such text. (4) To the best of the authors' knowledge, this is the first work on text localization in 3D video.

The remainder of the paper is organized as follows. Section 2 critically discusses related work. The proposed model is described in detail in Section 3, while Section 4 presents experimental results to validate the steps of the model. Finally, Section 5 concludes the paper and discusses future work.

II. Related Work

A number of methods have been developed in the past for text localization in 2D natural scene images and 2D video but there do not appear to have been any approaches proposed in the literature for text localization in 3D video. Therefore, the review below focuses on those most closely related methods of natural scene and video text localization.

A. Text Localization in Natural Scene Images

For text localization in natural scene images, powerful CNN based methods have been developed, which can be classified as *regression / anchor-based* methods, *segmentation-based* methods, and *hybrid* methods [11].

a. Regression/Anchor-Based Methods

The development of these approaches has been inspired by object detection, which considers the whole text as an object for text localization in natural scene images.

Liu et al. [12] proposed a method for Fast Oriented Text Spotting (FOTS) in natural scene images. The method focuses on a unified approach, which involves detection and recognition of text for achieving better spotting results. Liu et al. [13] proposed robust curved text localization in natural scene images based on conditional spatial expansion. He et al. [14] proposed a method for multi-oriented and multi-lingual text localization in natural scene images based on direct regression. Cheng et al. [15] proposed a direct regression scene text detector for the natural scene images based on positive-sensitive segmentation. Overall, when faced with the presence of shadows and decorative features associated with text in video or scene images, the regression-based methods miss the actual text information in order to predict the boundary points of the text. This is due to a lack of local information for predicting points. Therefore, the performance of this type of method is significantly reduced for text localization in 3D video.

The key issue of regression-based methods is the use of a rigid reference (anchor) for text localization in natural scene images. Due to this constraint, regression-based methods report poor results for images of irregularly sized and arbitrarily-oriented text. To alleviate this problem, anchor-based methods have been proposed to enhance text localization performance.

Deng et al. [16] proposed a real-time scene text detector using learner anchors. The method detects the location of an object of any shape by using learned proposals. Sheng et al. [17] proposed a single shot-oriented scene text detector with learnable anchors. Hou et al. [18] proposed a method for scene text localization using a hidden anchor mechanism. In that method, the predictions of anchors are considered as hidden layers and the weighted sum of the predictions is integrated into a direct regression-based network for text localization.

An overarching issue with the above methods is that if they define incorrect anchors for the text because of confusion between text and shadow pixels in 3D video or scene images, there is a high probability of considering a non-text region as text for predicting subsequent anchors. Therefore, it can be reasonably concluded that these methods will not be as effective for localizing text in 3D video.

b. Segmentation-Based Methods

The regression and anchor-based methods are not robust and accurate for localizing curved and short text in natural scene images. The motivation therefore arose to propose segmentation-based methods, which extract the information at pixel and character levels. Since those methods focus on pixel and character levels, they can be more robust to arbitrary orientation, short text, and irregular-sized text.

Baek et al. [8] proposed a method called Character Region Awareness for Text Detection (CRAFT). This method exploits the information at character level and the affinity between characters. Liao et al. [9] proposed a method for text localization in natural scene images based on the Differentiable Binarization (DB) concept. Based on Progressive Scale Expansion Network (PSENet) Wang et al. [10] proposed a method for text localization in natural scene images. The approach generates different kernels according to the scale of characters and text. Tang et al. [19] proposed a method for localizing dense and arbitrary oriented text lines in natural scene images. Their approach employs Instance Aware Component Grouping (ICG), which works based on a bottom-up procedure. Liu et al. [20] proposed a mask tightness text detector for arbitrarily shaped scene text localization. The model predicts a mask at pixel level to find text region based on learning and predefined knowledge. Dai et al. [21] proposed a method for curved text localization in natural scene images based on multi-scale context-aware feature aggregation. Zhang et al. [22] proposed a method for arbitrarily oriented text localization in natural scene images by exploring an omni-directional pyramid mask proposal network. Xing et al. [23] proposed a method based on a convolutional character network for text detection in natural scene images. Liu et al. [24] proposed a method based on context attention and repulsive text border. That method exploits context information extracted locally and globally for achieving better results. Zhang et al. [25] proposed a deep relational reasoning graph network for arbitrary shape text localization in natural scene images. Cao et al. [26] introduced a two-stage segmentation-based detector and it is coined as NASK (Need A Second look) for addressing challenges of arbitrarily shaped text detection in natural scene images. Cheng et al. [27] proposed position-sensitive segmentation-based model for text detection in natural scene images.

Overall, when images contain text with shadow and/or text is affected by perspective distortion due to different shooting angle, extracting features that represent a character is challenging. Moreover, not infrequently, the pixels that represent shadow share the properties of characters and then the segmentation-based methods fail to extract actual text information and hence their performance diminishes.

c. Hybrid Methods

Although segmentation-based methods can be more robust in the presence of challenges, the performance of deep learning models depends heavily on the number of samples and requires a large number of parameters. These hard constraints limit the generalization ability of those methods. To find a solution to this problem, hybrid methods have been proposed that integrate the merits of pixel/character level information (by extracting handcrafted features) and deep learning models.

Wang et al. [28] proposed a hybrid method, which combines respective advantages of segmentation and regression-based methods to overcome the limitations of either method on their own. Roy et al. [3] proposed text localization from multi-views of the scene images based on the Delaunay Triangulation concept. Xue et al. [29] proposed a method for arbitrarily oriented text localization in low-light natural images. The model integrates features extracted from the spatial and the frequency domains for enhancing low contrast text pixels in the images. Nag et al. [30] proposed a unified method for localizing text in images of marathon runners and sports players. The method combines dominant information detected by a handcrafted feature and a deep learning model to reduce false positives.

It can be reasonably argued that, since these hybrid methods involve handcrafted features for detecting local information of text regions, the features may not be adequate to differentiate text and shadow pixels in 3D video or 3D scenes. The review of the methods in this section indicates that existing models are capable of addressing the challenges of arbitrarily oriented text, different shaped text, and low light text detection in the natural scene images. However, none of those methods addresses the challenges encountered in text localization in 3D videos. This fact motivated the authors to propose the method for text localization in 3D video, described in this paper.

B. Text Localization in Video

Fassold and Germi [31] proposed a method for video text tracking in a real-time environment. Their approach combines deep learning and an object detector to achieve improved results. Raghunandan et al. [6] proposed a method for text localization in video. That approach uses bit plane slicing for detecting text in images/frames. Rasheed et al. [32] proposed a deep learning-based method for text localization in video frames. Shivakumara et al. [33] proposed a method for multi-oriented text localization in video images based on Fractal theory. The fractal concept is employed to enhance the low contrast text in video images. Wang et al. [5] proposed text localization and tracking in video using fully convolutional neural networks. Wang et al. [34] proposed text localization and tracking based on hybrid deep text detection and a layout constraint (text trajectories). Their approach combines object detection and semantic segmentation in a hybrid way for text

candidate detection in each frame. Yu et al. [35] proposed a method for text localization and tracking in video using deep learning models. That approach employs ConvLSTM to capture spatial structure information and motion memory.

Zhou et al. [36] proposed text localization in video using a YOLO deep learning model. The Efficient Convolutional Operators (ECO) method is used for tracking the text in the video. Local and global stitching is proposed to obtain a text panorama. Song et al. [37] proposed a Siamese network for video text frame detection. The Siamese network includes one branch for text similarity estimation and one for text identification. Wang et al. [38] proposed a method for scene text localization and tracking in video based on background cues. Their method employs background cues in identifying candidate text regions and uses spatial, shape and motional correlations between text and its background region for localization and tracking of that text. Cheng et al. [39] proposed a method for fast video text spotting based on text localization in video frames.

It can be noted from the above reviews on text localization in video that those methods focus on temporal information for text localization and tracking but none of the methods are applied to 3D video. Moreover, none of the methods addresses the challenges posed by shadow information in the video. However, there have been some attempts at the *classification* of 2D and 3D text in a video in order to hopefully improve subsequent text localization. For example, Xu et al. [40] proposed a method for the classification of 2D and 3D text regions in video. Their approach uses information on gradient direction for differentiating between pixels which represent 2D and 3D. Similarly, Zhong et al. [7] proposed a method for shadow detection in the 3D video to pave the way for 3D text recognition. Nandanwar et al. [41] proposed a method based on detecting common points for the classification of 2D and 3D text in video/scene images. Nevertheless, the main objective of the above three methods is to classify 2D and 3D text, not text localization. Recently, Chowdhury et al. [4] explored episodic learning based network for text detection on human body in sports images. However, the method is confined to video images of sports and the method works well for the images that contain text with cloth information.

In conclusion, it is observed that there is no robust method for text localization in the 3D video. Hence, this paper proposes a new method for text localization in 3D video by introducing a combination of novel concepts such as the Generalized Gradient Vector Flow (GGVF), the Wavefront, and the Adaptive B-Spline Polygon Curve Fitting Network (ABS-Net). These concepts enable to proposed method to successfully deal with the challenges of text localization in 2D and 3D video.

III. The Proposed Approach

The proposed approach comprises four key steps, namely, dominant point detection using GGVF, candidate point detection using wavefront, connected component analysis for constructing text patches, and using the ABS-Net for detecting and describing the text in the video. These key steps can be seen in the block diagram in Fig. 2. It can be observed that the stroke width (the thickness of a character stroke) usually has an almost constant value for each individual character, irrespective of 2D or 3D, in both video and scene images [42]. Moreover, the arrows of the Gradient Vector Flow (GVF), which are pointing towards the edge pixels, have opposite directions for the pixels representing the stroke width, exhibiting what is called an Opposite Direction Symmetry (ODS). Based on these observations, the proposed method employs Generalized Gradient Vector Flow (GGVF) [43] for defining ODS rather than using conventional GVF. If the pixels satisfy the ODS, they are considered as dominant points. It is therefore expected that this step generates dominant points, which represent text regardless of 2D and 3D texts.

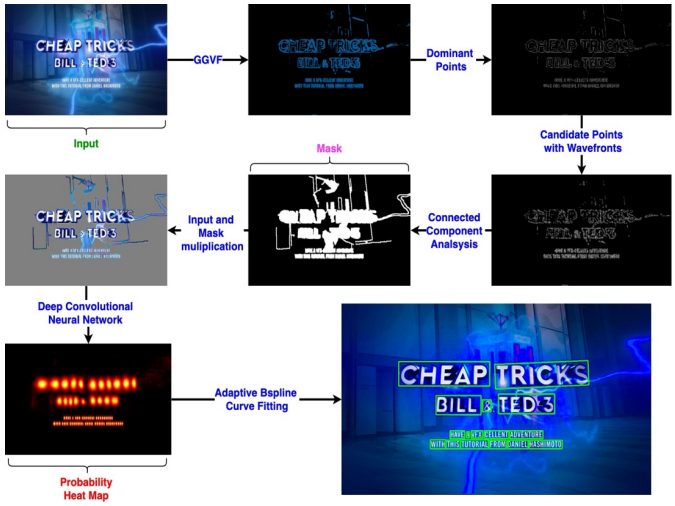
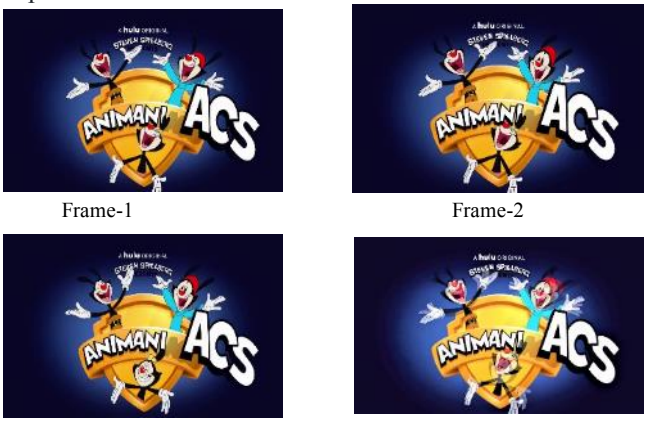


Figure 2. Block diagram of the proposed approach.

Due to the complexity of the problem, sometimes, the dominant point detection step misclassifies non-text pixels as text pixels. To overcome this problem, we introduce a novel concept called Wavefront, which predicts the dominant point based on the neighbor points according to the direction of the speed motion [44]. This observation inspired us to explore the same wavefront technique for predicting the values using a point's 8-neighborhood with the help of horizontal and vertical wavefronts. For each dominant point, if it is a text (or non-text) pixel, the Wavefront predicts the values using its 8-neighbors such that the predicted values exhibit a regular pattern. This results in robust candidate points by eliminating false dominant points. The proposed method then constructs text patches using candidate points with the help of Canny edge components and morphological operations. Finally, in order to describe text with accurate and tight bounding boxes (excluding excessive background pixels), we introduce a new idea of exploring Adaptive B-Spline Polygon Curve Fitting (ABS-Net) with a new deep learning model. The reason to explore B-Spline curve fitting is that it fits smooth curves using control points determined by local information, unlike Bezier curve fitting, which requires global information. Therefore, in contrast to conventional polygonal curve fitting, the ABS-Net is capable of overcoming the challenges posed by decorative text, which has irregularly shaped characters with non-uniform spacing and color bleeding.

It should be noted that, unlike a single-stage architecture system, which lacks robustness, adaptability and generalization

ability, the proposed multi-stage system is characterized by its flexibility, adaptability and generalization ability for successfully addressing the complex issues considered in this work. In addition, it is straightforward to expand and extend the multi-stage system for diverse data and new applications in contrast to single-stage system, which is not flexible for expansion and extension.



Frame-1 Frame-2

Frame-3 Average of frame-1, 2 and 3

Figure 3. Exploiting temporal information for enhancing the fine details in the images.

Finally, a decision has been made to use as input to the proposed approach the average of the first three successive frames out of the 25-30 frames per each second of video. This is due to a number of reasons. First. the average operation can be considered as a type is a type of a low pass filter, which attenuates high frequencies (e.g. corresponding to noise), resulting in enhanced images. Samples of three consecutive frames and average frame are shown in Fig. 3. Second, experimental observations indicate that the result of averaging the first three frames is almost indistinguishable from the average of all 25-30 temporal frames in each second of video, in terms of text localization performance. Third, averaging over only three frames significantly reduces the number of computations, ensuring the proposed method is practical. If video is not available, the proposed approach considers an individual image for text localization.

The following sections describe the main steps of the proposed method in detail.

A. Generalized Gradient Vector Flow (GGVF) for Dominant Points Detection

Inspired by the special property of Gradient Vector Flow (GVF), namely that the arrows of GVF pointing towards edges due to high force at edges [42], we exploit the same property for detecting dominant edge points in the average frame of the video. This helps us to reduce the complexity of the problem by removing unwanted background information in the input frame. The conventional GVF has some limitations, however, such as confusions in pointing arrows at corners and sometimes failing to point an arrow when the edge pixels suffer from degradations, low contrast and low resolution. These limitations motivated the authors to introduce the Generalized Gradient Vector Flow (GGVF), which is invariant to distortion, degradations to some extent, and provides correct arrows at corners of the edges. The formulation to derive GGVF from the conventional GVF is as follows.

The energy function of the GVF field $\boldsymbol{z}(x,y) = (u(x,y), v(x,y))$ is defined as

$$E = \iint g(|\nabla f|)(|\nabla u|^2 + |\nabla v|^2) + h(|\nabla f|)(\boldsymbol{z} - \nabla f)\, dx\, dy \quad (1)$$

$$\text{with } g(|\nabla f|) = e^{\frac{-\nabla f}{k}} \quad (2)$$

$$h(|\nabla f|) = 1 - g(|\nabla f|) \quad (3)$$

The first term in Equation (1) is a smoothing term, which produces a vector field. The second term is the data fidelity term, that drives the vector field v close to the gradient of the image i.e. ∇f. Parameter k acts as a weighing parameter that balances the smoothing and data fidelity terms. The value of k is related to the noise level. The higher the level of noise, the larger the value of k should be.

The energy function of the GGVF snake model is obtained by modifying Equation (1) as follows

$$E = \iint g(|\nabla f|)(\emptyset(|\nabla u|) + \emptyset|\nabla v|) + h(|\nabla f|)(\boldsymbol{z} - \nabla f)\, dx\, dy \quad (4)$$

where $\emptyset(|\nabla v|) = \sqrt{1 + |\nabla v|^2}$ and $\emptyset(|\nabla u|) = \sqrt{1 + |\nabla u|^2}$, and the definitions of $g(|\nabla f|)$and $h(|\nabla f|)$ remain same as in Equation (2) and Equation (3).

The solution to the energy function stated in Equation (4) is obtained using the calculus of variation and it is as follows:

$$\frac{\partial v}{\partial t} = g(|\nabla v|)\sqrt{1 + (|\nabla v|)^2}\nabla.\left(\frac{\nabla v}{\sqrt{1+(|\nabla v|)^2}}\right) - h(|\nabla v|)(v - \nabla f) \quad (5)$$

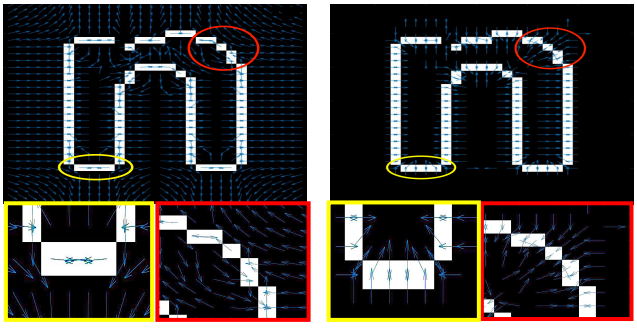
$$\frac{\partial u}{\partial t} = g(|\nabla u|)\sqrt{1 + (|\nabla u|)^2}\nabla.\left(\frac{\nabla u}{\sqrt{1+(|\nabla u|)^2}}\right) - h(|\nabla u|)(u - \nabla f) \quad (6)$$

The terms $\frac{\partial v}{\partial t}$ and $\frac{\partial u}{\partial t}$ can be described using the forward difference scheme. Hence, we can re-write Equation (5) and Equation (6) as

$$\frac{v^{n+1} - v^n}{\Delta t} = g(|\nabla v|)\left[\frac{(1 + v_y^2).v_{xx} + (1 + v_x^2).v_{yy} - 2v_x v_y v_{xy}}{1 + (|\nabla v|)^2}\right] - h(|\nabla v|)(v - \nabla f) \quad (7)$$

$$\frac{u^{n+1} - u^n}{\Delta t} = g(|\nabla u|)\left[\frac{(1 + u_y^2).u_{xx} + (1 + u_x^2).u_{yy} - 2u_x u_y u_{xy}}{1 + (|\nabla u|)^2}\right] - h(|\nabla u|)(u - \nabla f) \quad (8)$$

For our calculations we use standard values of all variables: $k = 3.8, \Delta t = 0.1$ *and* $t = 20$ [43].



(a) GVF (b) GGVF

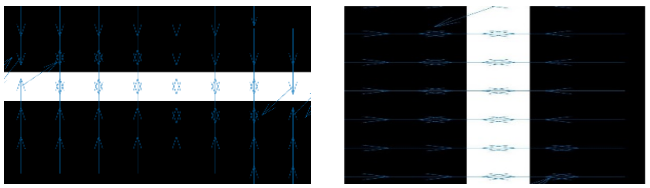
Figure 4. Comparison between conventional GVF and the proposed GGVF.

The noticeable difference in using GGVF over GVF is the use of the modified function $\emptyset(|\nabla u|) + \emptyset|\nabla v|$ defined in Equation (4). The result of including this function is the capturing of the sharp corners of the images with properly

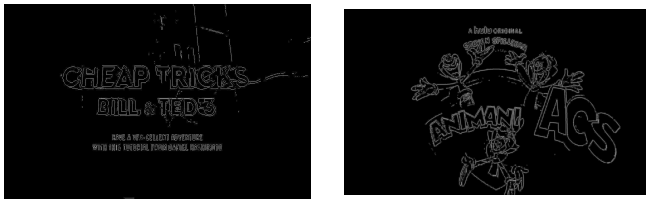
aligned gradient vectors as shown in Fig. 4, where it can be seen that GVF does not perform well at the corners, where opposite arrow direction for the edges would have been expected. This is because GVF is sensitive to low contrast and low-resolution edges. In the case of Fig. 4(b), it can be seen that GGVF overcomes the limitations of GVF. As mentioned earlier, if there is a pair of pixels representing the stroke width (the thickness of the stroke), the arrows should have an opposite direction, which is ODS. The edge pixels marked in the yellow and red color boxes shown in Fig. 4(a) do not satisfy ODS, while ODS is satisfied in the case of GGVF shown in Fig.4(b) (in Fig. 4, enlarged versions of the two marked portions of the upper part of the figures are shown in their respective lower parts). Furthermore, it is observed from the direction of arrows in both Fig. 4(a) and Fig. 4(b) that GGVF provides orderly and cleanly directed arrows while GVF does not. This fact enables the proposed method (using GGVF) to obtain accurate results.

The effects of GGVF on the input images of the example 2D and 3D video frames are shown in Fig. 5, where (a) is the result of GGVF showing all edge pixels for which arrows are present, (b) illustrates the ODS for the 2D and 3D video frames, (c) is the result of the edge pixels that satisfy ODS, referred to as *dominant points detection*. It is observed from Fig. 5(c) that most of the pixels which represent text are retained for both 2D and 3D video frames. It is also true that dominant pixel detection includes some pixels of non-text edges. This is justifiable at this stage because there are edges of background objects which satisfy ODS.

(a) GGVF for both 2D and 3D video frames.



(b) Opposite direction symmetry illustration



(c) Dominant points detection for 2D and 3D video frames.
Figure 5. The GGVF for dominant point detection.

B. Mini-Wavefronts for Candidate Points Detection

When the center pixel in 8-connectivity neighborhoods is representing text, one can expect almost the same properties among its neighbor pixels, such as color, direction, speed of motion, etc [42]. In the previous section, the proposed method uses a gradient direction-based feature for separating text and non-text pixels in the frame. In this section, we explore the direction of the speed of the motion through the Wavefront concept to improve the results of the dominant point detection step. The introduction of the Wavefront concept is inspired by the method proposed in [44] where horizontal and vertical mini wavefronts are used to predict the dominant point, through interpolation based on the direction of the speed of the motion. It is the authors' understanding that the nature of the predicted value depends on its neighbors' values. The formal steps for defining the Wavefront to predict the values in 8-connectivity neighborhoods are as follows.

For each dominant pixel in the frame, the proposed method defines a 3×3 window and these are considered as 8-connectivity grids. The mini-Wavefronts, namely the vertical and horizontal Wavefronts are illustrated in Fig. 6(a), where the center pixel is considered as the center node and other vertically (or horizontally) adjacent points are considered as two boundary points of Wavefronts. The step can be formulated as follows. Let $S(p) = \{(p, T_c);\ (p_l, T_l);\ (p_r, T_r)\}$ be the mini Wavefront core information, where (p, T_c) is the position and the tentative value in the mini Wavefronts center with (p_l, T_l) and (p_r, T_r) being mini wavefront boundaries, respectively. With these values, we define $T_{S(p)}(p_n)$ as the tentative value in p_n if the solution is propagated from the wavefront section $S(p)$. For each mini-Wavefront, the initial tentative value $(T_c / T_l / T_r)$ is 0. The tentative value $T_{s(p)}(p_s)$ for reaching point p_s can be estimated by approximating the integral between the reaching point and the Wavefront section in the direction of the motion. In other words, the process estimates the traveling cost using the speed of motion (F) given at p_s as defined in Equation (9).

The speed of motion F= (F_x, F_y) in an image with intensity I can be calculated as:

$$F_x(p) = \frac{\cos(|\nabla I|)}{1 + |\nabla I|^2}$$

$$F_y(p) = \frac{\sin(|\nabla I|)}{1 + |\nabla I|^2} \tag{9}$$

Thus, we can derive Equation (10) and Equation (11) for each of the mini wavefronts using Equation (9):

$$T^l_{s(p)}(p_s) = \min_{0<t<1} T(p^l_t) + ||p_s - p^l_t|| \frac{(F(p^l_t) + F(p_s)}{2} \tag{10}$$

where $p^l_t = (1-t)p + tp_l$ and

$$T^r_{s(p)}(p_s) = \min_{0<t<1} T(p^r_t) + ||p_s - p^r_t|| \frac{(F(p^r_t) + F(p_s)}{2} \tag{11}$$

where $p^r_t = (1-t)p + tp_r$

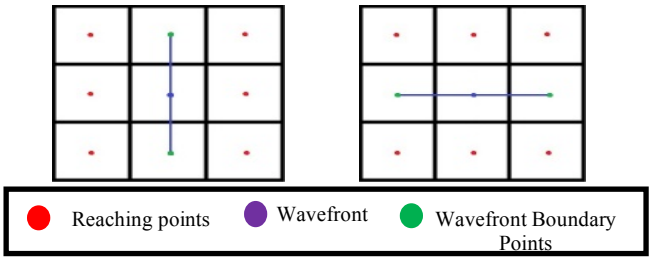
Finally,

$$T_{s(p)}(p_s) = \min(T^l_{s(p)}(p_s), T^r_{s(p)}(p_s)) \tag{12}$$

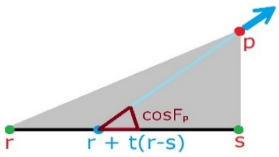
In Equation (12), by choosing an appropriate value for *t,* one can compute its respective $T(p_t)$ and $F(p_t)$ values. To achieve this, the proposed method uses gradient information for computing t values and linear interpolation to obtain the respective $T(p_t)$ and $F(p_t)$ values. The gradient information uses the direction of the speed of the motion to obtain the t values as shown in Fig.6(b), where it can be seen that the proposed method selects the intersection point between the

Wavefront segment and the line starting in p which follows the direction of the speed of motion. Let p be the node which is to be updated and predicted and r, s the endpoints of a wavefront section, as explained in Fig. 6(a). With $d = s - r$, the value of t is computed as defined in Equation (13).

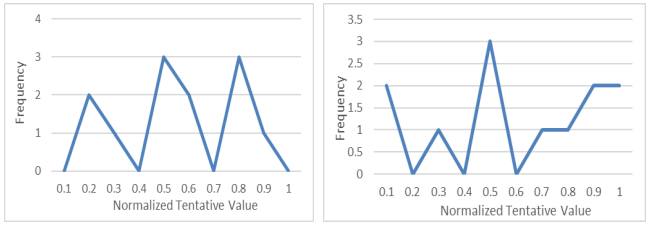
$$t = \max\left(min\left(\frac{F_y(r_x - p_x) - F_x(r_y - p_y)}{F_x d_y - F_y d_x}, 1\right), 0\right) \quad (13)$$



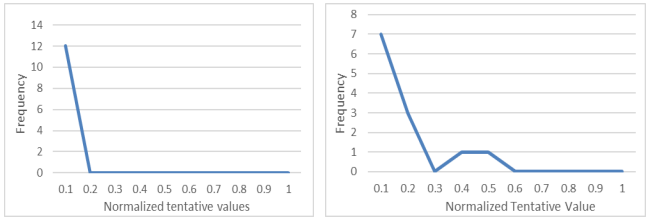
(a) Vertical and horizontal mini wavefronts for 8 neighbors



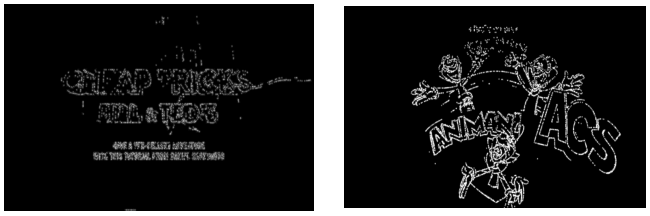
(b) Gradient method for *t* value estimation, *r* and *s* are endpoints, *p* is the point to be updated, where *Fp* is the Speed of motion at *p*. The value of *t* must be bounded between 0 and 1.



(c) Regular pattern of distribution for text dominant points



(d) Irregular pattern distribution for non- text dominant points.



(e) Candidate points for both 2D and 3D video frames.

Figure 6. Wavefront for candidate points detection.

To obtain the values for the $T(p_t)$, linear interpolation is proposed as defined in Equation (14) and Equation (15) with Wavefront centered at $p_{i+1,j+1}$. In the same way, the values of $F(p_t)$can be computed through the same interpolation as defined in Equation (14) and as shown in Equation (15).

$$T^l_{S(p_{i+1,j+1})}(p_t) = (1-t)T_{i+1,j+1} + tT_{i+1,j}$$

$$T^r_{S(p_{i+1,j+1})}(p_t) = (1-t)T_{i+1,j+1} + tT_{i,j+1} \quad (14)$$

Similarly,

$$F^l(p_t) = (1-t)T_{i+1,j+1} + tT_{i+1,j}$$

$$F^r(p_t) = (1-t)T_{i+1,j+1} + tT_{i,j+1}$$

$$F(p_t) = \min\left(F^l(p_t), F^r(p_t)\right) \quad (15)$$

For each dominant point in the image, the proposed method obtains 12 $T_{s(p)}(p_s)$ values which include 6 values from horizontal and 6 from vertical Wavefronts. It is expected that these values should exhibit a regular pattern if the pixel is text else an irregular pattern will be expected among those 12 values. To determine this observation, the proposed method draws a histogram for the 12 values as shown in Fig. 6(c) and Fig. 6(d), where a regular distribution can be seen for the dominant text pixels (Fig.6(c)) while an irregular distribution for non-text dominant pixels (Fig. 6(d)). If the dominant points satisfy this regular pattern distribution, the dominant pixels are considered as text candidate points else the proposed method discards the misleading dominant points. We use the Jarque-Bera Normal distribution test to determine whether the distribution is regular or irregular.

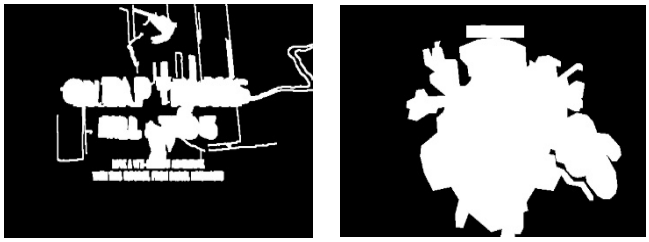
The effect of applying the Wavefront step can be noted in Fig. 6(e) for 2D and 3D frames, where it can be seen that a significant number of non-text pixels have been removed compared to the results in Fig. 5(c). In addition, the brightness of edge pixels has been increased in Fig. 6(e) compared to Fig.5(c). At the same time, we can see a reduction in the number of candidate pixels in Fig. 6(e) for both the images. This makes it more efficient to differentiate candidate points (which represent text) from non-text. The candidate points are considered in the next step of the proposed method: the construction of text patches.

C. Adaptive B-Spline Polygon Curve Fitting for Text Localization (ABS-Net)

It is noted that the deep learning models are more effective when regions are considered as input, rather than individual points. To exploit this advantage, the proposed approach reconstructs text patches using points representing text. For this purpose, the proposed method uses connected component analysis, which extracts edge components from the Canny edge image of the input images corresponding to each text representative. This results in an image with edge components.

To combine all the edge components in a single component, the proposed approach performs a morphological operation over edge component images, which fills the small (usually two/three pixels) gap between the edge components and then the closed contours are selected, accordingly. This results in text patches as shown in Fig. 7(a), where the whole patch can be seen as a single component, and it is considered as a text region. The actual color information in the input images corresponding to pixels of text patches is restored by extracting it from the input images as shown in Fig. 7(b). These regions are fed to the deep learning model to predict the character region heat map. The complete deep learning architecture is illustrated in Fig. 8 and its description in terms of learning and training are as follows. The first layer is a custom convolution

operation which is a filter (K) as defined in Equation (16) for enhancing fine details in the image [45].



(a) Constructing text patches for 3D frames.

(b) Text patches with color information for 3D frames.

Figure 7. Constructing text patches using connected component analysis.

$$K = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 10 & -1 \\ -1 & -1 & -1 \end{bmatrix} \quad (16)$$

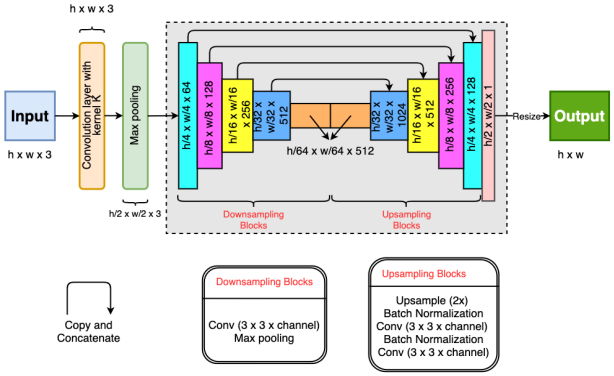


Figure 8. Proposed Deep Convolution neural network architecture

The second layer is a max pooling layer with kernel size 2 and stride 2. Then the proposed method uses VGG-16-Unet [46] with batch normalization as its backbone, which has skip connections in the decoding part similar to U-net [47] which aggregates low level features. For training the proposed model at character level, the character-level SynthText dataset is used in a weakly supervised manner, which is publicly available [48]. Transfer learning is used onto the proposed text localization model utilizing the pre-trained weights of CRAFT [8] text detection method in the proposed VGG-16-Unet backbone.

The final output of the proposed model is a single channel output, showing the probability of the character center with a Gaussian heat map. The instructions given in [8] are followed for training the model based on weak supervision. As mentioned above, the pre-trained weights of the CRAFT model are trained on the SynthText dataset [48] for 50k iterations, each benchmark dataset is then trained on to fine-tune the proposed model. The ADAM [49] optimizer is used in all training processes. For training and testing on image frames, the dataset (see Section IV) is split in a 4:1 ratio. For training in each epoch data augmentation techniques are also used randomly such as zoom in/out, Scaling, Cropping, Padding, Rotation, Affine transformation, which increase the dataset size to 4 times the actual training data size. The testing time of the proposed method is calculated as 168 ms for each image frame at HD resolution on an NVIDIA 2080 TI GPU.

For a word-level annotated sample w of the training data, let $R(w)$ and $l(w)$ be the bounding box region and the word length of the sample w, respectively. The character splitting process defined in [6] is used which provides the estimated character bounding boxes and their corresponding lengths of characters $l^c(w)$. Then the confidence score $s_{conf}(w)$ for the sample w is computed as defined in Equation (17).

$$s_{conf}(w) = \frac{l(w) - \min\left(l(w), l(w)\right) - l^c(w)}{l(w)} \quad (17)$$

and the pixel-wise confidence map S_c for an image is computed as defined in Equation (18)

$$S_c(p) = \begin{cases} s_{conf}(w) & p \in R(w) \\ 1 & otherwise \end{cases} \quad (18)$$

where p denotes the pixel in the region R(w). Our Loss function L is defined as in Equation (19).

$$L = \sum_p S_c(p) \times \left(\|S_r(p) - S_r^*(p)\|^2\right) \quad (19)$$

where $S_r(p)$ denotes the ground truth region score i.e 1, and $S_r^*(p)$ denotes the predicted region score. The $s_{conf}(w)$ increases as the proposed model is trained and, after the training, regions with $s_{conf}(w) < 0.5$ are disregarded as they have very low probability of containing text.

For the Gaussian probability heat map produced by the deep learning model as shown in Fig. 9(a) for 2D and 3D video frames, in order to fit accurate and tight bounding boxes for arbitrary oriented text lines, we explore B-Spline polygon curve fitting. It is inspired by the method in [50] where Bezier curve fitting is explored for constructing bounding boxes. Bezier curve fitting, however, does not work well for irregular sized text which contains different fonts and sizes of characters in the same line. In contrast, B-Spline curve fitting has the ability to fit smooth curves based on control points determined from local information [51] as shown in Fig. 9(b) where smooth and accurate bounding boxes can be seen fitted on arbitrarily oriented text lines.

To find control points, the proposed method draws a line through the local maxima of each character, which is then considered as the control line of the whole text line as shown in Fig. 9(b). Then the proposed method draws lines, perpendicular to the points of the control line, which are based on local maxima and extend towards the character boundaries in both directions, as shown in Fig 9(b). If the control line is a straight line, a conventional quadrilateral is used for fitting the bounding box, otherwise a B-Spline Polygon curve is used. The boundary points of characters are considered as control points as described in Fig 9(b). These control points are fed to the B-

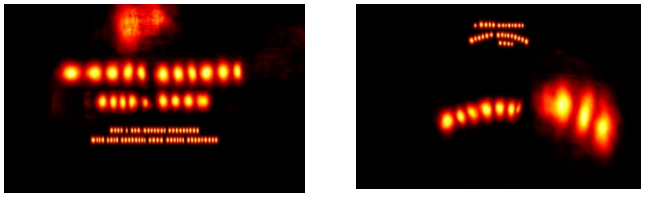
spline curve fitting process, which outputs the n-vertices polygon. In this work, according to experiments, a maximum feasible value of n is 100. The steps for polygon construction are presented in Equation (20). Let the B-spline Polygon fitting function be S(x), defined as in Equation (20):

$$S(x) = \sum_{j=0}^{n-1} c_j B_{j;j;t}(x) \quad (20)$$

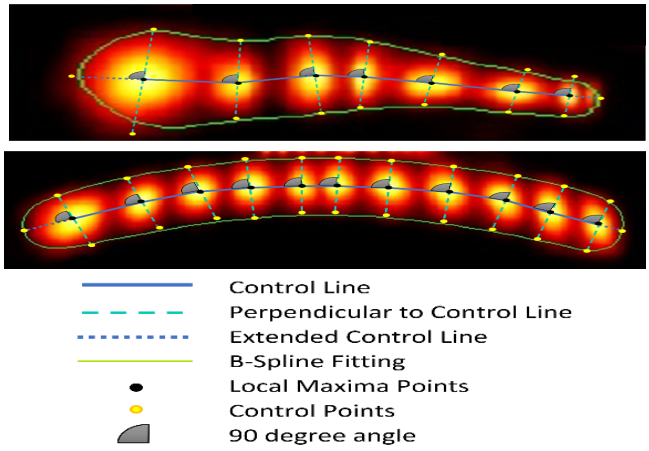
$$B_{i,0}(x) = 1, if\ t_i \le x \le t_{i+1}, otherwise\ 0$$

$$B_{i,k}(x) = \frac{x-t_i}{t_{i+1}-t_i} B_{i,k+1}(x) + \frac{t_{i+k+1}-x}{t_{i+k+1}-t_{i+1}} B_{i+1,k-1}(x)$$

Here, t is number of knots, c is the spline coefficients or control points, k is the B-spline order and n is the number of control points (c). Here $t \in \{-k, -k+1, \dots\dots, n+2k-1\}$ and $k = 2$.



(a) Predicted character region heat map for both 2D (left) and 3D (right) images



(b) Sample Adaptive B-Spline polygon curve fitting for fixing bounding boxes on curved text (for illustration purpose only)

(c). Text detection in both 2D and 3D frames.

Figure 9. Adaptive B-spline curve fitting Network for arbitrary oriented text detection

Due to the complexity of the problem, involving both 2D and 3D text in video and natural scene images, there is still a possibility of the proposed method misclassifying non-text regions as text. Therefore, to improve the performance of text localization, a combination of EAST [52], which detects text, and Tesseract [53], which recognizes text, is used to eliminate false positives. Text regions with bounding boxes are fed to the combination of EAST and Tesseract. If at least one of the characters is recognized, then the proposed method retains its text designation, otherwise it is discarded as a false positive. The text localization results of the proposed method for 2D and 3D frames are illustrated in Fig. 9(c), where it can be seen that the proposed method has correctly localized text in different orientations, in different font sizes as well as text composed of irregular characters.

IV. Experimental Results

To validate the proposed method, a new dataset was constructed for experimentation since there is no standard dataset for text localization in 3D video. The dataset includes text in arbitrary orientations, irregularly shaped text and irregularly sized text. Sources included 3D movies, sports and other internet video sources, such as YouTube, all contributing to a total of 400 videos captured at 25-30 frames per second. As result, the dataset provides 1200 frames including three consecutive frames for experimentation. Since the main objective of the proposed work is to localize text in 3D video as well as natural scene images, rather than just exploring temporal information, the proposed method only extracts three consecutive frames from the respective video. Each frame is converted to a standardized dimension of 1080×1920 pixels for experimentation. The frames were annotated manually at the word level according to the instructions given for the ICDAR dataset construction [54]. The dataset is divided into 80% and 20% for training and testing, respectively.

To evaluate and benchmark the ability of the proposed method to localize text in 2D video, the community standard ICDAR2013 Robust Reading Competition [51] dataset was used, which provides 28 videos with 15277 frames containing 2D text. That dataset also provides ground truth at the word level for all the frames. Noted that since the aim of the proposed work is to detect text in the video and not for tracking the text in the video, the metrics used in [51] are not used for evaluation, instead, standard text detection measures are used in this paper. Similarly, to evaluate the ability of the proposed method to localize text in 2D natural scene images, the following standard benchmark datasets (described in more detail further below) were used: SCUT-CTW1500, Total-Text, ICDAR 2019-ArT and DAST1500. For these experiments, the proposed method considers an individual image as input for text localization, without any temporal information.

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 CTW1500 dataset. However, most of the image contains English text lines. In this dataset, 1255 images for training and 300 images for testing are considered for experimentation. **ICDAR 2019 ArT**[55]: This dataset is a combination of the images of Total-Text, those of the CTW1500 dataset and those of the Baidu Curved Scene Text, which was created for detecting arbitrary-shaped text in the natural scene images. In total, the dataset contains 10,166 images, split into a training set of 5603 images and a testing set of 4563. **DAST1500**[19]: This dataset includes images with considers dense and arbitrarily shaped text. By dense it is meant that there exist several text lines in a single image without much space between them. This makes fixing

bounding box for arbitrary shaped text lines much more challenging. Out of a total of 1538 images, 1038 images are separated for training and 500 images for testing.

For evaluating the performance of the proposed method, the standard measures and evaluation scheme, Recall (R), Precision (P) and F-measure (F) have been used in all the experiments reported in this paper. The threshold for intersection-over-union (IoU) for classifying true and false positive is 0.5 according to the standard evaluation scheme [8-10, 12, 19].

For comparative study purposes, the proposed method is evaluated alongside the implementations of the most relevant prominent and recent existing methods available publicly for experimentation. These methods are: The Character Region Awareness for Text Detection (CRAFT) [8], the Differential Binarization Network (DB-Net) [9] and the Progressive Scalable Expansion Network (PSENet) [10]. These methods have been developed for text localization in natural scene images. As mentioned earlier in this paper, the reasons to consider the above methods for comparative study are that these approaches are the state-of-the-art methods and address several challenges which are very similar to text localization in 3D video. Another reason of considering those state-of-the-art methods is to evaluate in the first place how effective they are in text localization in 3D video. For the purpose of comparative study, the methods in [38, 39], developed for text localization and tracking in video using spatial-temporal information, have also been evaluated.

All existing methods were first trained using the ICDAR2015 dataset and the publicly available SynthText [48] dataset, following the instructions mentioned in [8]. Next, they were fine-tuned using the proposed 3D text dataset at the word level according to the instructions mentioned in each respective method. The details of training and testing for the proposed method are described in Section III.C.

A. Ablation Study

To achieve the best results for text localization in 3D video, the proposed method involves several key steps detailed earlier in this paper: dominant point detection using the GGVF approach, candidate point detection using the Wavefront approach, polygon boundary fitting for arbitrarily oriented text lines using adaptive B-Spline curve fitting, use of the combination of EAST and Tesseract as OCR for eliminating false positives, and use of the average frame for enhancing detail in the images/frames. To assess the contribution of each step in achieving the best text localization performance, we conduct the following experiments and calculate measures as reported in Table I on the proposed new dataset: Experiment-(i), where the proposed method considers the output of GGVF as input for text localization without candidate points. Experiment-(ii), where the proposed method replaces GGVF with GVF for text localization. Experiment-(iii), where the proposed approach does not consider dominant points for detecting candidate points, instead, it considers the Canny edge image of the input image. Experiment-(iv), where the proposed method replaces a B-Spline curve fitting with the conventional Bezier fitting approach used in [8]. Experiment-(v) is conducted to test the contribution of the combination of EAST + Tesseract used for eliminating false positives, where the proposed method calculates the measures without EAST + Tesseract. Finally, Experiment-(vi) assesses the contribution of the average frame, where the proposed method is evaluated on individual frames.

Table I. Effectiveness analysis of the key steps of the proposed method on our dataset based on different experiments.

Experiment #	Key Steps	P	R	F
(i)	Proposed using only the GGVF step	69.15	73.90	71.45
(ii)	Proposed using conventional GVF	68.31	70.79	69.53
(iii)	Proposed without GGVF	67.20	70.50	68.80
(iv)	Proposed using conventional Polygon Fitting [8]	52.69	79.26	63.30
(v)	Proposed without EAST + Tesseract	70.90	72.30	71.60
(vi)	Proposed on individual frames	67.82	74.11	70.83
Proposed Model		70.71	73.67	72.10

From experiments (i) and (ii) in Table I it can be seen that GGVF is better than GVF for localizing text in 3D video. When one compares the results of experiments (iii) and the proposed method reports poor results without Wavefront and GGVF, which indicates that GGVF and Wavefront significantly contribute in achieving the best results by the proposed method. The results of experiment (iv) and the proposed method show that the proposed ABS-Net is more effective than the Bezier curve fitting approach. Similarly, the results of experiment (v) and the proposed method show that the combination of EAST + Tesseract helps in improving the performance in terms of precision and F-measures. Finally, the results of experiment (vi) and the proposed method show that the average frame helps to improve text localization performance. Overall, it is noted from Table I that the key steps employed by the proposed method are effective in successfully addressing the challenges of text localization in 3D video.

B. Experiments for 3D Text Localization in 3D Video

Quantitative results from the proposed and existing methods are reported in Table II on the proposed new 3D video dataset. It is noted from Table II that the proposed method is the best in terms of Precision and F-measure compared to existing methods while the existing method in [10] is the best in terms of Recall compared to other methods, including the proposed method. When one compares the results of the three existing methods [8, 9, 10], the method in [8] is the best in terms of F-measure and the method in [10] is best in terms of Recall. This shows that the existing methods do not produce consistent results for text localization in 3D video. The main reason for the poor results of the existing methods is that the methods are not able to adequately cope with the challenges of shadows and decorative characters in video. As a result, the existing methods are prone to generating more false positives and hence Precision is low for all those methods compared to Recall.

Quantitative results from the proposed and existing methods for the benchmark video dataset (ICDAR2013) are reported in Table III. In this experiment, since the existing methods used individual frames for text localization, the proposed approach follows the same approach in calculating the evaluation measures for the purposes of comparative study. It is noted from Table III that the proposed method outperforms

the existing methods in terms of Recall, Precision, and F-measure. The reason for the inferior results of the existing methods is that the features used in those methods are sensitive to background complexity. On the other hand, since the method proposed in this paper employs dominant point detection by GGVF, candidate point detection by Wavefront, and Deep learning with adaptive B-spline curve fitting, all of which are invariant to the effects of 3D, 2D and multi-font text in video, it achieves better results compared to the existing methods. In addition, the use of temporal information (effectively the averaging process) for enhancing the details in the image/frame also contributes to the better performance.

To validate the above statement, the output of the GGVF and the Wavefront steps is fed separately into each of the three existing methods listed in Table II and Table III for text localization in video. Since the existing methods do not accept the dominant points given by GGVF or the candidate points given by the Wavefront step as input for text localization, the reconstructed text patches obtained by the connected component analysis step after the two respective steps are supplied to the existing methods to calculate the evaluation measures. It is observed from Table II and Table III that the results of all the existing methods improve when text patches are fed as input compared to existing methods applied directly on input images. However, this improvement is not sufficient to exceed the performance of the proposed method. One can therefore reasonably conclude that the GGVF and the Wavefront play a considerable role in improving the performance of text localization in 3D and 2D video or images.

Table II. Performance comparison of existing methods with the proposed method by considering original input images and the results of GGVF and Wavefront as input, using the proposed new dataset.

Methods	Original images			GGVF as input			Wavefront as input		
	P	R	F	P	R	F	P	R	F
PSENet [10]	62.6	**77.7**	69.3	62.81	78.18	69.65	64.71	78.15	70.80
DB-Net [9]	59.9	67.6	63.5	60.61	66.32	63.34	63.32	67.26	65.23
CRAFT [8]	65.6	77.0	70.8	66.46	77.51	71.56	66.10	77.94	71.53
Proposed model	**70.71**	73.67	**72.1**	--	--	--	--	--	--

Table III. Performance of the proposed and existing methods on benchmark dataset of video (ICDAR2013)

Methods	Original images			GGVF as input			Wavefront as input		
	P	R	F	P	R	F	P	R	F
Wang et al. [38]	58.34	51.74	54.45	61.5	56.9	59.1	63.5	60.8	62.1
Chen et al. [39]	**81.45**	60.23	69.25	80.7	63.3	70.9	81.1	64.3	71.7
Proposed Method	81.10	**64.60**	**71.90**	-	-	-	-	-	-

C. Experiments for 2D Text Localization in Natural Scene Benchmark Datasets

To demonstrate the effectiveness of the proposed method on text localization in 2D images, comparative experiments were conducted on four benchmark natural scene text datasets as discussed earlier. The results of the proposed and existing methods for the four datasets are reported in Table IV. It is observed from Table IV that the proposed method is the best in terms of Precision for all the four datasets compared to the existing methods. This indicates that the proposed approach is reliable in datasets of different complexities. When one examines the results of the proposed method, the method achieves almost consistent Precision for all the four datasets whereas the existing methods do not. This shows that the proposed method works well irrespective of the challenges represented in the different datasets. The reason for the inferior results of the existing methods is that the methods have their own inherent limitations. The Recall of the proposed method is low for all the four datasets compared to Precision. Besides, for the Total-Text dataset, the proposed method reports the lowest F-measure compared to the other datasets. The images of the Total-Text dataset are much more complex compared to other natural scene text datasets in terms of diversity. In this situation, the proposed GGVF and Wavefront steps may miss a minimal number of pixels that represent text information. Due to such a loss of pixels, the proposed method may miss text information and hence the recall can be low. This leads to lower F-measure for the Total-Text dataset. However, the precision achieved is the highest compared to existing methods for the Total-Text dataset. Overall, one can conclude that the proposed method is effective and useful.

(a) Proposed Method

(b) Existing DB-Net Method

(c) Existing CRAFT Method

Figure 10. Sample qualitative results of the proposed and existing methods for text localization in 3D and 2D video images.

In particular, when the results of text localization in 3D video and 2D images are compared, the proposed method reports better for localizing text in 2D images compared to 3D images. This is expected as the complexity of 2D images is lower than 3D videos. Therefore, one can conclude that the proposed method has the ability to achieve better results for 2D natural scene text images, without temporal information. The same conclusion can be drawn from the qualitative results of the proposed and existing methods shown in Fig. 10, where it can be seen that the proposed method localizes text well for the sample images from both the proposed new 3D video dataset and the benchmark 2D video images. However, despite the existing methods localize text well for 2D video images, they do not perform well for 3D video images as shown in Fig. 10. Furthermore, the results from the proposed method on the four benchmark 2D natural scene datasets shown in Fig. 11 prove that the proposed method localizes text well. Thus, the proposed method is independent of multiple font text and can cope well with the challenges of different datasets. Moreover, according

to the experimental results, one can also confirm that the proposed combination of feature extraction and deep learning is better than the methods which use only deep learning.

SCUT-CTW1500 (line level)

Total-text

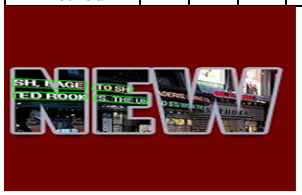
ICDAR ArT 2019.

DAST1500 (line level)

Figure 11. Sample text localization results of the proposed method on different benchmark datasets. (Note: Except DAST1500 and SCUT-CTW all datasets have word level annotations)

Table IV. Performances of the proposed and the existing methods on SCUT-CTW1500, Total-Text, ICDAR 2019 ArT and DAST1500 datasets.

Methods	SCUT-CTW1500			Total text			ICDAR 2019 Art			DAST1500		
	P	R	F	P	R	F	P	R	F	P	R	F
PSENet [10]	79.7	84.8	82.2	84.0	77.9	80.8	81.1	57.5	67.3	74.6	50.1	60.0
DB-Net [9]	80.2	**86.9**	83.4	87.1	**82.5**	**84.7**	56.0	**69.9**	62.2	51.3	55.1	53.2
CRAFT [8]	86.0	81.1	**83.5**	87.6	79.9	83.6	79.4	66.6	72.4	61.9	**88.2**	72.8
Xue et al. [29]	56.0	51.0	53.0	67.0	43.0	52.3	---	--	---	---	---	--
Dai et al. [21]	85.7	85.1	85.4	84.6	78.6	81.5	---	---	----	---	--	---
Liu et al. [20]	79.7	79.0	79.4	79.1	74.5	76.7	---	---	---	--	---	---
Proposed method	**92.7**	74.4	82.4	**91.1**	52.5	66.6	**90.4**	61.6	**73.3**	**86.5**	64.4	**73.8**



(a) Superimposed text

(b) Text and background same

Figure 12. Examples of limitations of the proposed model

Finally, it should be noted that sometimes, in superimposed text (when a piece of text is written over another piece of text) as shown in Fig. 12(a), the proposed system does not work well. In this case, the local information of the inner text is missed by the candidate point detection step. Similarly, when the color of the text is similar to the background, as shown in Fig. 12(b), the proposed system misses some points during the dominant point detection step and hence the subsequent steps fail to recover the missing text. Therefore, there is a scope for improvement in the future to handle such challenges. One possible way to address such challenges is to design a new unified architecture for robust dominant point detection and candidate point detection by exploring temporal information as well.

V. Conclusion and Future Work

This paper has proposed a new approach for text localization in 3D video. The proposed approach is the first to address the challenges of this complex application domain. It exploits the Generalized Gradient Vector Flow (GGVF), which defines opposite arrow direction symmetry for detecting dominant pixels irrespective of 2D and 3D text in images/frames. To improve the results of GGVF which uses only the gradient direction, the proposed method introduces a novel concept called Wavefront approach for eliminating false dominant points by considering the direction and speed of motion for text pixels. The candidate points are used to form text patches based on connected component analysis. Next, a deep learning model, integrating adaptive B-Spline curve fitting, has been proposed for the final text localization using very accurately fitting bounding boxes/polygons. Experimental results on a 3D video dataset and four benchmark 2D natural scene text datasets show that the proposed method outperforms existing methods. However, in special case such as superimposed text and text in similar colour to the background, as well as when the image is affected by blur and in too low a resolution, the performance of the proposed method degrades. Therefore, future work will focus on exploring temporal information in an effective way to overcome those limitations of the proposed approach.

Acknowledgement

The authors of this work received support from the Natural Science Foundation of China under Grant 61672273, and partial support from FRGS grant (FP104-2020), Ministry of Higher Education, Malaysia. This work is also partly supported by TIH, ISI, Kolkata.

References

[1] Y. Peng and J. Chi, “Unsupervised cross-media retrieval using domain adaptation with scene graph”, IEEE Trans. CSVT, pp 4368-4379, 2020.

[2] Z. Zhang, D. Xu, W. Ouyang and C. Tan, “Show, tell and summarize: Dense video captioning using visual cue aided sentence summarization”, IEEE Trans. CSVT, pp 3130-3139, 2020.

[3] 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, 129, pp 92-100, 2020.

[4] P. N. Chowdhury, P. Shivakumara, R. Raghavendra, S. Nag, U. Pal, T. Lu and D. Lopresti, “A new episodic learning based network for text detection on human body in sports images”, IEEE Trans. CSVT, 2021.

[5] Y. Wang, L. Wang, F. Su and J. Shi, “Video text detection w1ith fully convolutional network and tracking”, In Proc. ICME, pp 1738-1743, 2019.

[6] K. S. Raghunandan, P. Shivakumara, S. Roy, G. H. Kumar, U. Pal and T. Lu, “Multi-script-oriented text detection and recognition in video/scene/born digital images”, IEEE Trans. CSVT, 29, pp 1145-1162, 2019.

[7] W. Zhong, A. N. J. Raj, P. Shivakumara, Z. Zhuang, T. Lu and U. Pal, “A New Shadow Detection and Depth Removal Method for 3D Text Recognition in Scene Images”, In Proc. ICIMT, pp 277-281, 2018.

[8] Y. Baek, B. Lee, D. Han, S. Yun and H. Lee, “Character region awareness for text detection”, In Proc. CVPR, pp. 9365-9374, 2019.

[9] M. Liao, Z. Wan, C. Yao, K. Chen and X. Bai, “Real-time scene text detection with differentiable binarization”, In Proc. Proc. AAAI, 2020.

[10] W. Wang, E. Xie, X. Li, W. Hou, T. Lu, G. Yu, S. Shao, “Shape Robust Text Detection With Progressive Scale Expansion Network,”, in Proc. CVPR, pp 9328-93337, 2019.

[11] S. Long, X. He and C. Yao, “Scene text detection and recognition: The deep learning era”, IJCV, 2020.

[12] 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.

[13] 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.

[14] 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.

[15] P. Cheng, Y. Cai and W. Wang, “A direct regression scene text detector with positive-sensitive segmentation”, IEEE Trans. CSVT, pp 4171-4181, 2020.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] S. Zhang, Y. Liu, L. Jin, Z. Wei and C. Shen, “OPMP: An omni-directional pyramid mask proposal network for arbitrary-shape scene text detection”, IEEE Trans. MM, 2020.

[23] L. Xing, Z. Tian, W. Huang and M. R. Scott, “Convolutional character networks”, In Proc. ICCV, pp 9126-9136, 2019.

[24] 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.

[25] 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.

[26] M. Cao, C. Zhang, D. Yang and Y. Zou, “All you need is a second look: Towards arbitrarily-shaped text detection”, IEEE Trans. CSVT, 2021.

[27] P. Cheng, Y. Cai and W. Wang, “A direct regression scene text detector with position-sensitive segmentation”, IEEE Trans. CSVT, pp 4171-4181, 2020.

[28] 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.

[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] H. Fassold and R. Ghermi, “OminiTrack: Real time detection and tracking of objects, text and logos in video”, In Proc. ISM, pp 245-246, 2019.

[32] J. Rasheed, A. Jamil, H. B. Dogru, S. Tilki and M. Yesiltepe, “A deep learning based method for Turkish text detection from videos”, In Proc. ELECO, pp 935-939, 2019.

[33] P. Shivakumara, L. Wu, T. Lu. C. L. Tan, M. Blumenstein and B. S. Anami, “Fractals based multi-oriented text detection system for recognition in mobile video images”, Pattern Recognition, 68, pp 158-174, 2017.

[34] X. Wang, X. Feng and Z. Xia, “Scene video text tracking based on hybrid deep text detection and layout constraint”, Neurocomputing, 363, pp 223-235, 2019.

[35] H. Yu, C. Zhang, X. Li, J. Han, E. Ding and L. Wang, “An end to end video text detector with online tracking”, In Proc. ICDAR, 601-606, 2019.

[36] T. Zhou, K. Wang, J. Wu and R. Li, “Video text processing method based on image stitching”, In Proc. ICIVC, pp 561-566, 2019.

[37] H. Song, H. Wang, S. Huang, P. Xu, S. Huang and Q. Ju, “Text Siamese network for video textural key frame detection”, In Proc. ICDAR, pp 442-447, 2019.

[38] L. Wang, Y. Wang, S. Shan, and F. Su, “Scene Text Detection and Tracking in Video with Background Cues”, In Proc. ICMR, pp 160–168, 2018.

[39] Z. Cheng, J. Lu, Y. Niu, S. Pu, F. Wu, and S. Zhou. "You Only Recognize Once: Towards Fast Video Text Spotting”, In Proc. ACMMM, pp 855-863, 2019.

[40] 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.

[41] L. Nadanwar, P. Shivakumara, A. Kumar, T. Lu, U. Pal and D. Lopresti, “A new common points detection method for classification of 2D and 3D texts in video/scene images”, In Proc. DAS, 2021.

[42] V. Khare, P. Shivakumara, C. S. Chan, T. Lu, L. K. Meng, H. H. Woon and M. Blumenstein, “A novel character segmentation-reconstruction approach for license plate recognition”, ESWA, pp 219-239, 2019.

[43] S. Zhu and R. Gao, “A novel generalized gradient vector flow snake model using minimal surface and component-normalized method for medical image segmentation”, In Proc. BSPC, pp 1-10, 2016.

[44] B. Cancela, M. Ortega and M G. Penedo, “A Wavefront Marching Method for Solving the Eikonal Equation on Cartesian Grids”, In Proc. ICCV, pp 1832-1840, 2015.

[45] S. S. Somvanshi, P. Kumar, S. Tomar and M. Singh, “Comparative statistical analysis for the image enhancement techniques”, International Journal of Image and Data Fusion, pp 131-151, 2017.

[46] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition”, In Proc. ICLR, 2015.

[47] O. Ronneberger, P. Fischer, and T. Brox. U-net, “Convolutional networks for biomedical image segmentation”, In Proc. MICCAI, pages 234–241, 2015.

[48] A. Gupta, A. Vedaldi, and A. Zisserman, “Synthetic data fortext localisation in natural images”, In Proc. CVPR, pp 2315–2324, 2016.

[49] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization”, In Proc. ICLR, 2015.

[50] Y. Liu, H. Chen, C. Shen, T. He, L. Jin and L. Wang, “ABCNet: Real time scene text spotting with adaptive Bezier curve Network”, In Proc. CVPR, pp 9809-9818, 2020.

[51] M. Unser, A. Aldroubi and M. Eden, “B-spline signal processing. I. Theory”, IEEE Trans. Image Processing, 41, pp821-833, 1993.

[52] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He and J. Liang, “EAST: an efficient and accurate scene text detector”, In Proc. CVPR, pp. 5551-5560, 2017.

[53] J. Singh and B. Bhushan, "Real Time Indian License Plate Detection using Deep Neural Networks and Optical Character Recognition using LSTM Tesseract," In Proc. ICCCIS, pp 347-352, 2019.

[54] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. Bigorda, S. R. Mestre, J. Mas, D. Mota, J. Almazan, and L. Heras”, “ICDAR 2013 robust reading competition”. In Proc. ICDAR, pp 1484–1493, 2013.

[55] C. K. Chng, Y. Liu, Y. Sun, C. C. Ng, C. Luo and Z. Ni, “ICDAR2019 robust reading challenge on arbitrarily-shaped text-RRC-ArT”, pp 1571-1576, 2019.

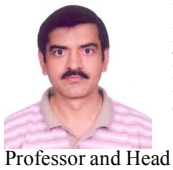
Lokesh Nandanwar is a pursuing Master in computer science at University of Malaya, Malaysia. He received Bachelor of Technology degree in Information Technology from Department of Computer Science and Engineering, NIT Durgapur, India in 2019. His research interests include deep learning, image processing, computer vision and video text analysis.

Palaiahnakote Shivakumara is an Associate Professor at 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.Sc., M.Sc., M.Sc Technology by research and Ph.D. degrees in computer science respectively in 1995, 1999, 2001 and 2005 from University of Mysore, Karnataka, India. He has been serving as Associate Editor for Pattern Recognition (PR), Springer Nature Computer Science (SNCS), ACM Transactions Asian and Low-Resource Language Information Processing (TALLIP). He received a prestigious award, "Dynamic Indian of the Millennium" from KG foundation, India for his research contribution to computer science field. He has published more than 200 papers in conference and journals. His research interests are in the area of image processing, Document image analysis and Video text processing.

Raghavendra Ramachandra received the bachelor's degree from the University of Mysore (UOM), Mysuru, India, the master's degree in electronics and communication from Visvesvaraya Technological University, Belgaum, India, and the Ph.D. degree in computer science and technology from the UOM and Institute Telecom, and Telecom Sudparis, Évry, France (carried out as a collaborative work). He was a Researcher with the Istituto Italiano di Tecnologia (IIT), Genoa, Italy. He is currently appointed as a Professor with the Norwegian Biometric Laboratory, Norwegian University of Science and Technology (NTNU), Gjøvik, Norway. He has authored several articles. His main research interests include statistical pattern recognition, applied machine leaning, deep learning, data fusion schemes, and random optimization, with applications to biometrics, image/video processing. multimodal biometric fusion, human behavior analysis, and crowd behavior analysis. He has been a Speaker at international conferences.

Tong Lu received the Ph.D. 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 is also a member of the National Key Laboratory of Novel Software Technology in China. He has published over 130 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.



Umapada Pal received his Ph.D. from Indian Statistical Institute. He did his Post-Doctoral at INRIA (Institut National de Recherché en Informatique et en Automatique), France. From January 1997, he is a Faculty member of the Computer Vision and Pattern Recognition Unit, Indian Statistical Institute(ISI), Kolkata and at present he is a Professor and Head of the Computer Vision and Pattern recognition Unit of ISI. Because of his significant impact in the Document Analysis research domain of Indian language, TC-10 and TC-11 committees of IAPR (International Association for Pattern Recognition) presented 'ICDAR Outstanding Young Researcher Award' to Dr. Pal in 2003. He is the Editorial board member several journals like PR, IJDAR, PRL, ACM Transactions on Asian Language Information Processing, IET-Biometrics etc. He is the Co-Editor-in-Chief of the Springer Nature Computer Science journal. He is a fellow of IAPR and Senior member of IEEE.

Apostolos Antonacopoulos leads the Pattern Recognition and Image Analysis (PRImA) research lab at the School of Science, Engineering and Environment at the University of Salford, UK where he currently holds the post of Professor of Pattern Recognition. He received his PhD from the University of Manchester, Institute of Science and Technology (UMIST), UK in 1995. In 2005, he received the IAPR/ICDAR Young Investigator Award for "Outstanding service to the ICDAR community and his innovative research in historical document processing applications". Professor Antonacopoulos has worked and published extensively on various problems in Document Analysis and Understanding (Image Enhancement, Segmentation, Recognition, Performance Evaluation) as well as on other applications of Pattern Recognition and Image Analysis. He currently serves on the Executive Committee of the International Association for Pattern Recognition (IAPR) as Past President, having also previously held the posts of President, Treasurer, 1st and 2nd Vice President. He has also chaired or served as a member of a number of IAPR and other international professional committees.

Dr. Yue Lu is a Professor of the Department of Computer Science and Technology, East China Normal University, and is presently serving as the Director of Shanghai Key Laboratory of Multidimensional Information Processing. He received his B.S. degree in wireless technology and M.S. degree in telecommunications and electronic system, both from Zhejiang University in 1990 and 1993 respectively, and his Ph.D. degree in pattern recognition and intelligent system from Shanghai Jiao Tong University in 2000. From 1993 to 2000, he was an engineer at the Third Research Institute of Posts and Telecommunications Ministry of China. Before he joined East China Normal University in 2004, he was a research fellow with the Department of Computer Science, National University of Singapore. In 2010, he was a visiting scientist at the Centre for Pattern Recognition and Machine Intelligence (CENPARMI), Concordia University of Canada for six months. His research interests include pattern recognition and machine learning, image processing and machine vision, biometrics, machine intelligence and intelligent system. Professor Lu has contributed to more than 120 reviewed publications in journals and conferences, and holds 17 authorized patents. He is serving as associate editors for the Pattern Recognition, and the International Journal of Pattern Recognition and Artificial Intelligence.