2022 26th International Conference on Pattern Recognition (ICPR)
August 21-25, 2022, Montréal, Québec, Canada

TWD: A New Deep E2E Model for Text Watermark/Caption and Scene Text Detection in Video

Ayan Banerjee[1], Palaiahnakote Shivakumara[2], Parikshit Acharya[1], Umapada Pal[1] and Josep Llados Canet[3]
[1]Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India.
[2]Facutlty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia.
[3]Computer Vision Center, Universitat Autònoma de Barcelona, Spain.
ab2141@cse.jgec.ac.in, shiva@um.edu.my, parikshitofficial@gmail.com, umapada@isical.ac.in, josep@cvc.uab.es

2022 26th International Conference on Pattern Recognition (ICPR) | 978-1-6654-9062-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICPR56361.2022.9956279

Abstract—**Text watermark detection in video images is challenging because text watermark characteristics are different from caption and scene texts in the video images. Developing a successful model for detecting text watermark, caption, and scene texts is an open challenge. This study aims at developing a new Deep End-to-End model for Text Watermark Detection (TWD), caption and scene text in video images. To standardize non-uniform contrast, quality, and resolution, we explore the U-Net3+ model for enhancing poor quality text without affecting high-quality text. Similarly, to address the challenges of arbitrary orientation, text shapes and complex background, we explore Stacked Hourglass Encoded Fourier Contour Embedding Network (SFCENet) by feeding the output of the U-Net3+ model as input. Furthermore, the proposed work integrates enhancement and detection models as an end-to-end model for detecting multi-type text in video images. To validate the proposed model, we create our own dataset (named TW-866), which provides video images containing text watermark, caption (subtitles), as well as scene text. The proposed model is also evaluated on standard natural scene text detection datasets, namely, ICDAR 2019 MLT, CTW1500, Total-Text, and DAST1500. The results show that the proposed method outperforms the existing methods. This is the first work on text watermark detection in video images to the best of our knowledge.**

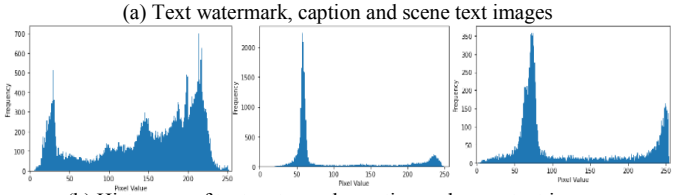
Keywords—Deep learning, U-Net, FCENet, Scene text detection, Video text detection, Watermark text detection.

I. INTRODUCTION

There is a tremendous progress on text detection in the natural scene and video by addressing several open challenges according to surveillance, monitoring, and retrieval applications [1, 2]. In spite of a large number of methods for text detection in natural scenes and videos, the researchers ignored text watermark, which is part of most video content to protect through copyright [3]. In the same way of applications mentioned above for text detection, text watermark detection is essential for multimedia content protection and security applications. When the image (keyframe) extracted from the video contains three types of text, namely, text watermark, caption (edited text), and scene text (natural text), the methods developed in the past may not work well because the scope of the existing methods is limited to caption and scene text in video images but not text watermark. In addition, the nature and characteristics of each type are different. There are different types of watermarking, such as covertly, which is not visible clearly, imperceptible, visible, and fragile, which is used to tamper with the original data [3]. However, this work considers images extracted from a video containing covertly and imperceptible watermark for detection as these are widely used in the case of copyright of the video.

For example, scene text usually has a complex background because it is wild, caption text (subtitles) has a homogeneous background because it is edited on a homogeneous background, while text watermark has low visibility compared to image content and partially mixes with the background. The same characteristics can be seen in Fig. 1, where the histograms are drawn for the pixel values of text watermark, caption and scene text images, respectively in Fig. 1(a)-(b). It is noted from the histograms of different texts that scattered peaks and valleys for text watermark image, spike peak for caption text image, while wide peaks for the scene text image. Scattered peaks and valleys indicate there is no clear separation from text pixel and background pixels, Spike peaks indicate homogenous background with uniform text pixels, while wide peaks indicate non-uniform text and background pixels. This shows that although these three types are text, share different characteristics and nature. Hence, detecting text watermark, caption and scene text successfully under a single end-to-end framework is challenging compared to normal scene text detection. Therefore, the existing scene text detection methods [4, 5] are limited to one type of text.

(a) Text watermark, caption and scene text images



(b) Histogram of text watermark, caption and scene text images

Fig. 1. Different characteristics of text watermark, caption, and scene text images.

It is evident from the illustration presented in Fig. 2, where we can see the video image containing only text watermark (left column), video image containing text watermark, caption, and scene text (middle), and the image containing only scene text (right column). For these images, the existing method [5] which is the state-of-the-art method for text detection in complex document and natural scene images, detect scene text well but miss text watermark. On the other hand, the proposed TWD model detects text accurately, irrespective of the type of text and images. Although, the existing model [5] is trained with samples of text watermark images, the method does not perform well for text watermark images. This shows that due to different nature

and characteristics of text watermark images illustrated in Fig. 1, which may cause loss of shapes, structure of characters, increasing number of training samples do not work, rather it leads to problem of overfitting and generalization. This situation motivated us to enhance low contrast text of watermark images such that the structure and shape can be restored and share the common text properties with caption and scene texts. This is the rationale behind proposing a new end-to-end model, which includes U-Net3+ for enhancement and segmentation of text and FCENET for text detection in this paper.

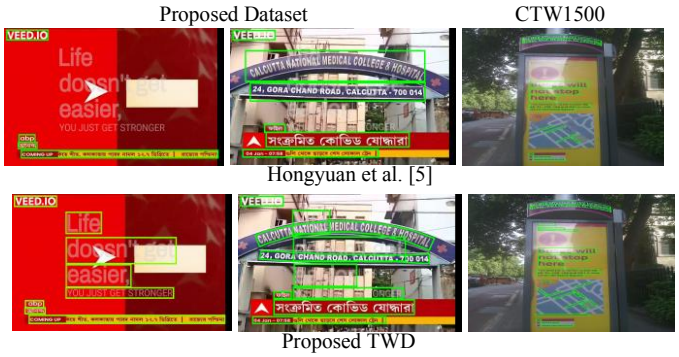


Fig. 2. Performance of the proposed and existing methods on the images containing text watermark, caption, and scene text. The proposed dataset provides text watermark and CTW1500 does not.

Since the input image extracted from the video contains three different types of texts, to standardize the quality and contrast, inspired by U-Net3+ which has used successfully for medical image segmentation [6], we propose a deep learning model based on U-Net3+, which enhances low contrast text (text watermark) without affecting high-quality text (caption and scene text). In the same way, to address challenges of arbitrary orientation shapes of the text, inspired by the FCENet model [7], which was developed to handle the above challenges of scene text detection, we explore FCENet by considering the enhanced image as input for successful text detection regardless of different types of text, which is named as a new SFCENet.

Thus, the key contribution of the proposed work is as follows. (i) Exploring U-Net3+ to enhance low-contrast and quality text without affecting high-quality text in the input images. (ii) Exploring SFCENet for successfully detecting text watermark, caption, and scene text. (iii) End-to-End model by integrating U-Net3+ and SFCENet for detection irrespective of text types.

The rest of the paper is organized as follows. The earlier methods developed for scent text detection and video text detection are reviewed in Section II. Section III presents U-Net3+ for image enhancement and SFCENet model for detection. Experimental results are discussed in Section IV. Lastly, the paper is concluded in Section V.

II. Related Work

Since use of watermark to declare the copyright ownership is not a new problem in the field of security, there are several models for protecting image and video through watermarking [8-11]. However, these methods focus on creating watermark for protecting copyright and ownership but not detecting and thus these methods are not suitable for text watermark in the images.

Similarly, there are several methods for text detection in natural scene images [12-23]. Most methods use deep learning models for addressing challenges of natural scene text detection. These methods are successful in achieving better results for text detection because the methods consider the relationship between pixels, characters for feature extraction.

In summary, although the models work well for the image with a complex background, in the case of video images containing both caption and scene texts, one cannot expect the same performance. This is because of superimposed text (caption text), which alters the content of textual information and hence it affects the relationship between text pixels as well as background pixels. In the same way, the methods were developed in the past for detecting text in video [5, 24-27]. In summary, since these methods use either temporal information or enhancement steps to improve the quality of the video images, these models are not effective for still images. In addition, the methods are not consistent for detecting text in both video and still images.

However, it is noted from existing scene text and video text detection methods, none of the methods focus on text watermark in the video and still images. There are some methods for text watermark detection in document images [28, 29]. But these methods are limited to plain documents but not scene text images with complex backgrounds. Thus, we can conclude that detecting, text watermark, caption and scene text in video images is an open challenge. Therefore, this work aims at developing a new Deep End-to-End Model for Text Watermark/caption and scene detection (TWD).

III. Proposed Approach

This work considers key frames containing text watermark, caption text and scene text for detection, which is the same as a still image extracted from video. Therefore, the proposed work does not use temporal information for text detection. As illustrated in Fig. 1, detecting three types of text in the video images is challenging. Thus, the aim of the proposed work is to develop a generic approach for detecting three types of text successfully. The key challenge is that variation in contrast and resolution of the text for not achieving better results.

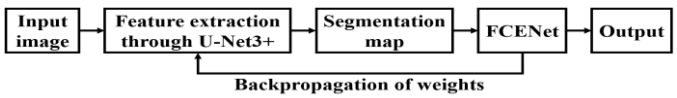


Fig. 3. Workflow of the proposed TWD

To addresses this challenge, inspired by the work in [30], where the U-Net and its subsequent networks are used for enhancing text in day and night images, we explore the U-Net in different way for enhancing low contrast text without affecting high contrast text such that structure and shape of the characters can be restored. Therefore, we introduce a new layered feature extraction network, U-Net3+ to segment each type of text in different layers of upsampling-downsampling. However, it is noted that for text detection, we need a network whose parameters are compatible with the U-Net3+. To alleviate this challenge, noted from [7] the FCENet [7] has the similar

properties with U-Net3+ as they perform stepwise Fourier embeddings for arbitrary text detection, we explore FCENet to detect text from the segmented results given by U-Net3+. Thus, the combination of U-Net3+ and FCENet is novel for text watermark/caption and scene detection in video images. Therefore, the proposed work is named Deep End-to-End Model for Text Watermark Detection (TWD) and its workflow is shown in Fig. 3.

A. U-Net3+ for Text Feature Extraction

The details of U-Net3+ for feature extraction from the input images containing text watermark, caption, and scene text are as follows. The proposed U-Net3+ has full-scale skip connections (i.e., the interconnection between encoder and decoder (IED) as well as intra connection between decoder layers (ICD)), which make it superior to other feature extraction networks (FEN) (U-Net, U-Net++, ResNet, etc.) as the dense connections of FEN are short of exploring significant information from full scales and unable to learn positional information of an object. Fig. 4 describes how an efficient feature map can be constructed with U-Net3+.

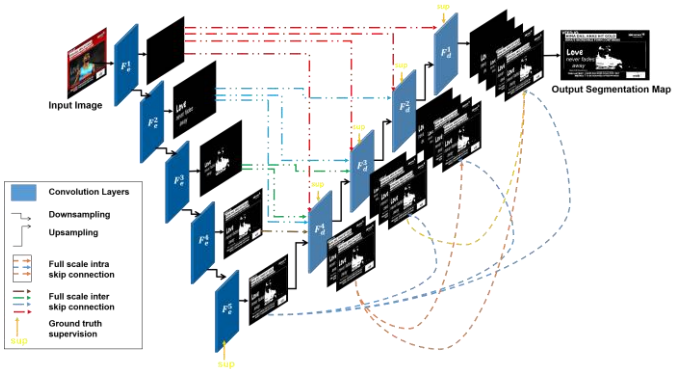


Fig. 4 Workflow of the U-Net3+ architecture.

It can be observed in Fig. 4 that the extracted feature maps from the same-scale encoder layer ($F_e^i; i = 1,2..,5$) are directly fed to the decoder layer through IED (represented in red, green, blue and brown color, respectively, in Fig.3). This set of IED is dedicated to delivering the low-level (i.e., watermark information) from the smaller scale encoder layer (F_e^1 *and* F_e^2) through non-overlapping max pooling. Similarly, the chain of ICD (represented in blue, yellow and orange respectively) transmits the high-resolution semantic information from the larger-scale decoder layer (F_d^4 *to* F_d^3) by bilinear interpolation. Now the five increasing (i.e., low to high contrast information) resolution feature maps are bound together with the quantity of channels to lessen the pointless data. It happened to us that the convolution with 64 channels of size 3×3 can be an astounding decision. To consistently consolidate the shallow impeccable data with profound semantic data, we further play out a total element instrument on the connected component map from five scales, which comprises 320 channels of size 3×3, a group standardization, and a ReLU actuation work. Officially, we figure out the skip associations as follows: let *i* list the down-testing layer along with the encoder, *N* (N=5 here) alludes to the complete number of the encoder. Now with this semantic information, the feature maps are stacked in every decoder layer (F_d^i) as they are superimposed to get a clearer view of all the semantic information of the textual region in a single feature map using Equation (1).

$$F_d^i = H(\sum_{i=1}^{N} C(D(F_e^i)) \cdot C(F_e^N) \cdot \sum_{i=1}^{N-1} C(U(F_d^i))) \quad (1)$$

Where C() represents the convolutional operation, H() indicates the convoluted feature aggregation mechanism with batch normalization and RELU activation, and D() and U() denote the down and up-sampling mechanism while (·) denotes the concatenation.

(a) Only watermark and subtitles

(b) Watermark, natural scene text and subtitles

(c) Only natural scene text: CTW1500 dataset.

Fig. 5. Layer-wise feature extraction with U-Net3+.

To further enhance the region of the watermarks to get the same resolution of all types of texts, a multi-scale structural similarity index (MS-SSIM) loss function is proposed to assign higher weights to the fuzzy region. Benefiting from it, the U-Net 3+ will keep an eye on fuzzy region as the lesser the regional distribution difference, the lower the MS-SSIM value. Two corresponding $N \times N$ sized segments are cropped from the segmentation result S and the ground truth mask G, which can be denoted as $s = \{s[:j = 1, \dots, N^2\}$ *and* $g = g[:j = 1, \dots, N^2\}$, respectively. The MSSSIM loss function of p and g is defined as in Equation (2).

$$L_M = 1 - \prod_{n=1}^{N} \left(\frac{2 \times \mu_{s_j} \times \mu_{g_j} + I_1}{\mu_{s_j}^2 + \mu_{g_j}^2 + I_1} \right)^{\beta_n} \times \left(\frac{2 \times \sigma_{s_j} \times \sigma_{g_j} + I_2}{\sigma_{s_j}^2 + \sigma_{g_j}^2 + I_2} \right)^{\gamma_n} \quad (2)$$

Here, $\mu_{s_j}, \mu_{g_j}, \sigma_{s_j}, \sigma_{g_j}$ are the mean and standard deviations of s and g; β_n and γ_n define the relative importance of two different types of text information will be changed depending on the pixel resolution of those types of texts. Two small constants $I_1 = 0.0001$ and $I_2 = 0.0009$ are used to tackle divided by zero error. By combining focal loss ($L_{F_d^i}$), MS-SSIM loss and IoU loss (L_{IoU}) [31], a hybrid loss for segmentation in three-level hierarchy – pixel-, patch- and map-level has been obtained, which is able to capture both high as well as low resolution text region with clear boundaries. The hybrid segmentation loss (L_F)is defined as in Equation (3).

$$L_F = \frac{L_{F_d^i} + L_M + L_{IoU}}{3} \quad (3)$$

This loss function is revised further to make the model compatible for end-to-end training.

The effect of the proposed U-Net3+ for the images shown in Fig. 1 is illustrated in Fig. 5(a)-(c), where one can see for each of different types of input images, as the layer increases, a contrast of the text improves without affecting background

information. Therefore, one can infer that the proposed framework can converge efficiently with fewer parameters.

It is noted from Fig. 5 that the extracted semantic information is performed from lower-to-higher contrast regions, and they are stacked as well as superimposed such that it highlights text irrespective of types. In this way, the proposed U-Net3+ reduces the effect of different types of text for improving text detection performance.

B. SFCENet for Text Watermark Detection

The enhanced text image shown in Fig. 5 given by U-Net3+ is the input for text detection. We propose Stacked Hourglass Encoded Fourier Contour Embedding Network (SFCENet) inspired from FCENet [7], where two stacked hourglass encoders (SHE) have been utilized to tackle the input image and the segmentation map such that it can fuse both the information for accurate detection as shown in Fig. 6.

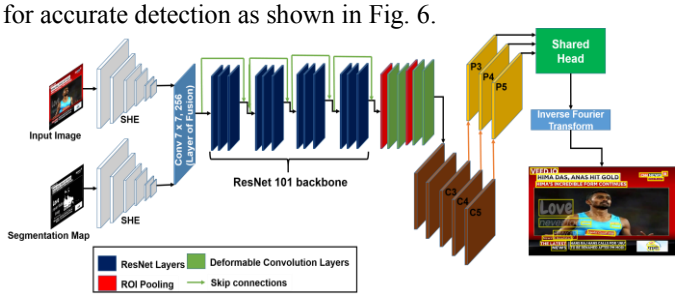


Fig. 6. SFCENet Architecture

Here the SHE [32] consists of convolutional layers to hold the positional information of the textual region in the input image as well as segmentation map, which will be passed through a fusion convolutional to fuse two positional vectors so that the rest of the network trained with the sufficient features of all kinds of texts. The objective function of the fusion layers is defined in Equation (4).

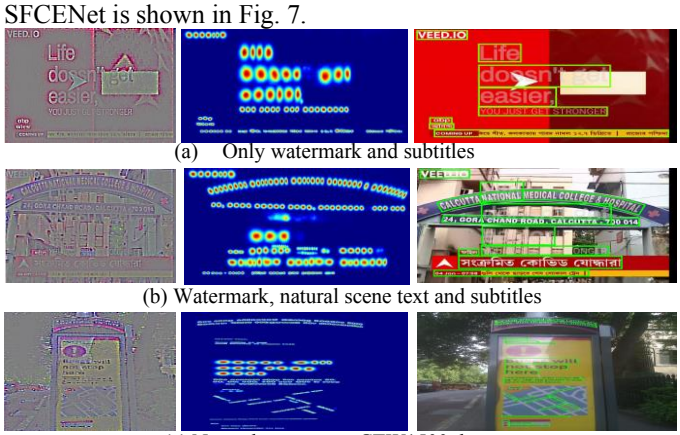
$$F_{con} = \alpha \cdot SHE_i^{\alpha}(z) \times \beta \cdot SHE_s^{\beta}(z) \quad (4)$$

where, F_{con} is the output feature vector; α and β are the hyperparameter of the two SHE to train the input image; SHE and segmentation map SHE are set to 0.000001 and 0.01. respectively; $SHE_i(z)$ and $SHE_s(z)$ represent the output feature vector image SHE and map SHE, respectively. Here, $\alpha<\beta$ as the segmentation map contains a more accurate textual position should be prioritized. To implement this layer, the activation function of the dedicated layer should be replaced by Equation (4). Next, this fused feature vector is fed to the backbone obtained with ResNet101 [33], Deformable Convolution Network [34], followed by a Feature Pyramid Network [35] to tackle the multi-contrast text information in a top-down scheme with inverse Fourier prediction header using Equation (5).

$$c_k = \frac{1}{n}\sum_{i=1}^{n} f\left(\frac{P_3}{C_3} \cdot \frac{P_4}{C_3} \cdot \frac{P_5}{C_5}\right) \times e^{-2\pi i} \quad (5)$$

where, n is the length of the obtained feature vector and f is the contour embedding function. This intermediate information passed through the Shared Head consists of two branches: classification and regression branches comprising three 3×3 convolutional layers and one 1 × 1 with Leaky-ReLU activation function.

In the classification head, we foresee the per-pixel veils of Text Regions (TR). We observe that the Text Center Region (TCR) forecast can additionally work on the exhibition. We accept this because it can successfully sift through low contrast pixels around the text region. Similarly, in the regression head, the Fourier mark vector of one text is relapsed for every conjugate pixel in the text. To manage text occasions of various scales, the elements of P3, P4 and P5 are answerable for low (watermark), medium (scene text), and high (subtitles) contrast text occurrences, individually. Later on, these two pieces of information are put into a spatial domain, and the final detection results are obtained through inverse Fourier Transform (IFT) followed by non-maximum suppression (NMS). The effect of SFCENet is shown in Fig. 7.



(a) Only watermark and subtitles

(b) Watermark, natural scene text and subtitles

(c) Natural scene text: CTW1500 dataset

Fig. 7. The effect of the proposed SFCENet for text detection.

It is observed from Fig. 7 that the proposed SFCENet detects text accurately regardless of text types. This makes sense because of the feature extraction through SHE and information fusion, which play a vital role in detecting the text with variable contrast. We minimize the reconstructed text contours to train the model, which is treated as a loss function as defined in Equation (6).

$$L_{SFCE} = \delta_1 L_{tr} + \delta_2 L_{tcr} + \frac{\delta_3}{n}\sum_{i=1}^{n} f^{-1}\left(\frac{c_i}{n}\right) \quad (6)$$

where, L_{tr} and L_{tcr} are the root mean square error loss (RMSE) at TR and TCR, respectively, and δ_1, δ_2 *and* δ_3 are the hyperparameters set to 1 to maintain the balance feature distribution.

C. End-to-End Model for Text Detection

In order to obtain an E2E model, the corresponding ground truth of each and every image plays a crucial role. Besides that, we propose a combinational to perform this end-to-end training. This end-to-end architecture is illustrated in Fig. 8. To train the TWD model, we need an adversarial loss function to minimize the loss through U-Net3+ as well as SFCENet and backpropagate it for accurate detection. The loss function of the SFCENet is defined in Equation (6). So, we obtain the revised loss function of the U-Net3+ as shown in Equation (7).

$$L_{U3} = 1 - \max\left(S_f \times M, I_i \times G, L_F\right) \quad (7)$$

Where, S_f is the final segmentation map, M is the ground truth mask, I_i is input image pixel matrix and G is the ground truth

of the corresponding image. This information obtains end-to-end training loss in Equation (8).

$$L_E = \emptyset \cdot L_{U3} + \varepsilon \Delta L_{SFCE} \quad (8)$$

Where, $\emptyset$ is the model parameter, ε is the learning rate, Δ is the differentiable operator to backpropagate the loss.

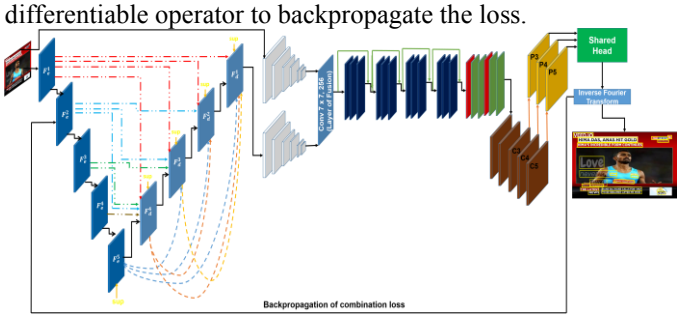


Fig. 8. End-to-End architecture of TWD

To optimize the proposed network a partially coupled non-linear parameter optimization algorithm (PCNPO) [36] has been utilized as it has the potential to minimize the adversarial loss of the multivariate hybrid models. We define the objective function of the PCNPO in Equation (9).

$$\theta_{L_E} = argmin_{\emptyset,\varepsilon}(F_d^i(L_E), F_{con}(L_E) \cdot c_k) \quad (9)$$

Where, θ_{L_E} is the optimized number of parameters. Now, the implementation details of the optimized TWD are obtained as follows.

During training, the input pixel matrix is standardized to the standard size of 512×512 with zero padding. Tests are directed on a PC with Intel i7-9700 CPU and NVIDIA RTX A6000. The whole model is carried out utilizing Tensorflow2. X. We use Adam Optimizer [37] with group size 7 and learning 10^{-6} rate where weight decay 7×10^{-4} is for 1000 epochs. The $\emptyset$ in Equation (8) is picked as 0.3 following our tests on the sample set. During training as well as testing, we consider 75% of the population to train the model and 25% of the samples for testing. The same setup is utilized for the analyses on all the datasets.

IV. Experimental Results

There was no standard dataset for evaluating the proposed method on text watermark/caption and scene text detection in the keyframes extracted from the video. We create our own dataset, which includes scene images of varieties of text watermark in terms of contrast, quality, color and shape variations, and different orientations and scripts. Most of the images in our dataset contain at least two types of texts: text watermark, caption, and scene texts. To test the objectiveness of the proposed method, we also consider the standard dataset of natural scene text detection, namely, ICDAR 2019 MLT [38], CTW1500 [39], Total-Text [40], and DAST1500 datasets [21], which are widely used for evaluating natural scene text detection methods, because the aim of the proposed work is to develop a model for not only detecting text watermark but also caption text in the video image and scene text in natural scene images. In the same way, to show the performance of the proposed method is effective, we consider the following state-of-the-art methods for comparative study [18, 19, 20, 22, 23]. The reason to choose the methods mentioned above for comparative study with the proposed method is that the approaches can handle challenges similar to text watermark video images. The existing methods are re-trained on our TW-866 dataset to tune the hyper parameters for comparative study with the proposed method. The same experimental set up has been used for both the proposed and existing methods for calculating measures.

A. Dataset Creation and Evaluation

The proposed TW-866 dataset consists of 866 images collected from 100 videos by sparse sampling with a sampling rate of 10^7. These videos are collected from various Bengali and English news, which contains text watermark, natural scene text as well as subtitles (caption texts). It is noted that all the frames contain watermarks, however, out of 866 images 572 images contain all three types of text and the rest of the images contain only watermark and subtitles. In the model, 606 images are used for training, 174 images for validation and 86 images for testing.

For evaluating the proposed and existing methods, we use standard measures called Pixel Accuracy (PA), which is defined as the ratio of a number of classified pixels and an actual number of pixels, Intersection over Union (IoU), which finds standard pixels between classified pixels by the proposed method and the actual number of pixels. Parameters used in the proposed networks are in terms of Million (M). For text detection performance, we use standard measures called Precision (P), Recall (R), and F-Measure (F). The evaluation scheme is used as per the instruction provided in the [31, 32].

B. Ablation Study

In our method, we propose U-Net3+ to enhance low-contrast text without affecting high-quality text. To assess the effectiveness of the proposed step, we conduct experiments listed in Table I with different versions of U-Net by calculating PA and IoU for our dataset, TW-866. In addition, we also consider the number of parameters used in the proposed U-Net3+ to show that the U-Net3+ is efficient compared to other versions, including baseline (U-Net). It is observed from Table I that as the version of U-Net improves, the PA and IoU also improve. However, this is not true for the parameter *M*. When we add residual and dense operation to U-Net, the number of parameters increases drastically compared to baseline (U-Net). On the other hand, the proposed U-Net3+ requires fewer parameters than all other versions. Therefore, from these experiments, we can infer that the proposed U-Net3+ is effective for detecting three types of text.

Similarly, for detection and designing end-to-end-model, we propose SFCENet for which the baseline architecture is ResNet and integrate U-Net3+ and SFCENet, respectively. To show the effectiveness for improving performance of the proposed model in terms of recall, precision and F-measure, and end-to-end, we conducted experiments listed in Table II with a different combination using our dataset. Table II shows that the proposed SFCENet and the combination of U-Net3+ and SFCENet are better than all other combinations, respectively. In addition, the results of end-to-end in Table II show that the text detection performance improves when we add more operations to baseline end-to-end architecture. Therefore, one can conclude that the

proposed end-to-end model is capable of handling three types of text in the video images.

Table I. Ablation study of feature extraction on TW-866 dataset

Methods	Pixel Accuracy (PA)	Intersection over Union (IoU)	Parameters (in M)
U-Net	0.894	0.915	22.7
U-Net++	0.913	0.928	21.9
Residual U-Net	0.924	0.946	29.3
Dense U-Net	0.947	0.937	32.1
U-Net3+	**0.958**	**0.962**	**17.6**

Table II. Ablation study for text detection and end-to-end performance on TW-866 dataset

Methods	P	R	F	Methods	P	R	F
ResNet101	78.8	78.1	78.3	U-Net3+ and ResNet101	81.7	81.2	81.5
ResNet101+DCN	81.3	78.8	79.1	U-Net3+ and ResNet101+DCN	83.4	81.7	82.3
ResNet101+DCN+FPN	82.8	82.1	82.4	U-Net3+ and ResNet101+DCN+FPN	85.7	85.2	85.5
ResNet101+DCN+FPN+SH	84.7	84.1	84.4	U-Net3+ and ResNet101+DCN+FPN+SH	86.8	86.3	86.7
ResNet101+DCN+FPN+SH+IFT	86.7	86.2	86.3	U-Net3+ and ResNet101+DCN+FPN+SH+IFT	89.6	89.1	89.4
ResNet101+DCN+FPN+SH+IFT+NMS	87.1	86.8	86.9	U-Net3+ and ResNet101+DCN+FPN+SH+IFT+NMS	89.2	88.9	89.0
SFCENet	**89.1**	**88.6**	**89.0**	U-Net3+ and SFCENet	**92.0**	**91.5**	**91.8**

Fig. 9. Text detection results of TWD on TW-866

C. End-to-End Experiments on our Text Watermark Dataset

Qualitative results of the proposed end-to-end model on TW-866 are shown in Fig. 9, where it is noted that the proposed model detects text watermark, caption text as well as scene text in the images, accurately. This shows that the proposed model generic for text detection The same conclusions can be drawn from quantitative results of the proposed and existing models reported in Table III, where it can be seen that the proposed end-to-end model achieves the best precision, recall, and F-measure compared to the existing models. The reason for the poor results of the existing methods is that the models were developed for scene text detection but not for text watermark and caption detection. On the other hand, since the proposed U-Net3+ and SFCENet effectively handle challenges of three types of text as noted from Table I-Table II, the proposed end-to-end model is the best.

Table III. Performance of TWD on TW-866

Wang et al. [22]			Li et al. [20]			RFRN [19]			Wu et al. [23]			Beak et al. [18]			TWD		
P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
85.6	85.2	85.4	84.8	86.3	85.3	85.1	87.5	86.2	88.3	88.1	88.1	84.3	89.8	86.9	**92.0**	**91.5**	**91.8**

D. End-to-End Experiments on Natural Scene Benchmark Dataset

Qualitative results of the proposed method on different benchmark scene text datasets are shown in Fig. 10, where the proposed end-to-end model detects text correctly irrespective of orientation, shape, script, and complex background. Therefore, we can assert that the proposed model is robust to different texts. To validate the robustness of the proposed model, the quantitative results of the proposed model are compared with the results of the existing models as reported in Table IV. It can be observed from the Table that the proposed model outperforms the existing methods for CTW1500 and ICDAR MLT 2019 datasets in terms of precision, recall and F-measure. However, for Total-Text and DSTA1500 datasets, the proposed model achieves the best recall and precision, respectively, compared to the existing methods. The models [19, 23] achieve the best precision and F-measure for Total-Text, respectively and also, the method [23] reports the best recall for the DSTA1500 dataset compared to other methods including the proposed method. This is because the models [19, 23] are robust techniques developed to tackle several challenges of scene text detection. However, the methods [19, 23] are not the best at recall for Total-Text and Precision and F-measure for DSTA1500 dataset compared to the proposed model.

CTW1500 DAST1500 Total-Text

Fig.10. Text detection on different benchmark scene text datasets with the proposed model.

Overall, when we compare the results on our dataset and benchmark scene text datasets, the proposed model outperforms the existing models in handling multiple types of texts. Hence, the proposed end-to-end model is effective and useful.

Table IV. Performance of the proposed and existing methods on CTW1500, Total-Text, DAST1500, and ICDAR 2019 MLT

Methods	CTW1500			Total-Text			DAST1500			ICDAR 2019 MLT		
	P	R	F	P	R	F	P	R	F	P	R	F
Wang [22]	85.6	84.7	85.1	88.5	88.9	88.7	84.8	85.2	84.9	88.5	89.8	88.7
Li [20]	86.9	84.9	86.2	87.9	85.2	86.6	85.9	81.1	83.7	89.5	80.3	84.4
RFRN [19]	87.9	86.8	87.2	**89.9**	82.7	86.9	86.3	89.1	88.2	89.6	82.3	86.4
Wu [23]	90.4	89.2	89.8	87.8	91.6	**90.1**	89.0	**90.9**	90.0	90.3	87.2	89.9
Beak [18]	86.0	81.1	83.5	87.6	79.9	83.6	83.6	68.2	73.9	88.2	78.2	82.9
TWD	**91.3**	**92.1**	**91.7**	86.9	**91.7**	89.8	**89.7**	90.7	**90.2**	**91.1**	**93.6**	**92.8**

V. Conclusion and Future Work

In this paper, we have proposed a new end-to-end deep model for text watermark/caption and scene text detection in the images extracted from video. To reduce the effect of contrast, quality, and color variations of different types of text, namely, text watermark, caption, and scene text, we have proposed U-Net3+ to enhance low-contrast text without affecting high-quality text. For detecting text of different types, we have proposed SFCENet to take care of challenges of scene text detection. Furthermore, the proposed work integrates U-Net3+ and SFCENet as an end-to-end model for detection. The results on our dataset and different benchmarks of scene text show that the proposed model is generic and robust to multiple types of texts compared to the existing methods. However, sometimes, when text watermark overlaps with caption and scene texts, the performance of the method degrades.

Acknowledgement

This project was funded by Ministry of Higher Education of Malaysia for the generous grant Fundamental Research Grant Scheme (FRGS) with code number FRGS/1/2020/ICT02/UM/02/4. The work received partial support from TIH, Indian Statistical Institute, Kolkata, India and also from the Spanish project RTI2018-095645-B-C21.

REFERENCES

[1] P. Shivakumara, S. Roy, H. A. Jalab, R. W. Ibrahim, U. Pal, V. Khare and A. W. B. A. Wahab, "Fractional means based method for multi-oriented keyword spotting in video/scene/license plate images", Expert Systems with Applications, pp 1-19, 2019.

[2] H. Mokayed, P. Shivakumara, H. H. Woon, M. Kankanhalli, T. Lu and U. Pal, "A new DCT-PCM method for license plate number detection I drone images", Pattern Recognition Letters, pp 45-53, 2021.

[3] R. Sripradha and K. Deepa, "A new fragile image-in-audio watermarking scheme for tamper detection", In Proc. ICISSS, pp 967-973, 2020.

[4] E. S. Jung, H.G. Son, K. Oh, Y. Yun, S. Kwon and M. S. Kim, "DUET: Detection Utilizing Enhancement for Text in Scanned or Captured Documents" , In Proc. ICPR, 2021.

[5] H. Yu, Y. Huang, L. Pi, C. Zhang, X. Li and L. Wang, "End-to-end vidoe text detection with online tracking", Pattern Recognitoin, 2021.

[6] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamaoto, X, Han, Y. W. Chen and J. Wu, "UNET3+: A full scale connected UNet for medical image segmentation", In Proc. ICASSP, pp 1055-1059, 2020.

[7] Y. Zhu, J. Chen, L. Liang, Z. Kuang, L. Jin and W. Zhang, "Fourier contour embedding for arbitrary-shaped text detection", In Proc. CVPR, pp 3123-3131, 2021.

[8] M. Gupta and R. Kishore, "A surevy of watermarking techniques using deep neural network architectures", In Proc. ICCCIS, pp 630-635, 2021.

[9] W. Qi, S. Guo and W. Hu, "Generic reversible visible watermarking via regularized graph Fourier transfrom coding", IEEE Trans. Image Processing, Vol. 31, 2022.

[10] L. Xiong, X. Han, C. N, Yang and Y. Q. Shi, "Robust reversible watermarking in encrypted image with secure multi-party based on lightweigth cryptography", IEEE Trans. Cicuits and Systems for Video Technology, Vol. 32, 2022.

[11] B. Ning, B. Niu, H. Gum, Y. Huang and S. Zhang, "Research and development of copyright regsitration and monitoring system based on digital watermarking and fingerprint technology", In Proc. ICCST, pp 354-358, 2021.

[12] P. Dai, H. Zhang and X. Cao, "Deep multi-scale context aware feature aggregation for curved scene text detection", IEEE Trans, MM, pp 1969-1984, 2020.

[13] P. Dai, H. Zhang and X. Cao, "Progressive contour regression for arbitrary-shape scene text detection", In Proc. CVPR, pp 7389-7398, 2021.

[14] Z. Hu, X. Wu and J. Yang, "TCATD: Text contour attention for scene text detection", In Proc. ICPR, pp 1083-1088, 2021.

[15] K. Yuan, D. He, X. Yang, Z. Tang, D. Kifer and C. L. Giles, "Follow The Curve" Arbitrarily oriented scene text detection using key points and curve prediction", In Proc. ICME, 2020.

[16] M. Zhao, W. Feng, F. Fin, X. Y. Zhang and C. L. Liu, "Mutually guided dual-task network for scene text detection", In Proc. ICPR, pp 6928-6934, 2021.

[17] A. Zhu, H. Du and S. Xiong, "Scene text detection with selected anchors", In Proc. ICPR, pp 6608-6615, 2021.

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

[19] G. Deng, Y. Ming and J. H. Xue, "RFRN: A recurrent feature refinement network for accurate and efficient scene text detection", Neurocomputing, pp 465-481, 2021.

[20] J. Li, Y. Lin, R. Liu, C. M. Ho and H. Shi, "RSCA: Real-time Segmentation-based Context-Aware Scene Text Detection", In Proc. CVPR, pp 2349-2358, 2021.

[21] 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, 106954, 2019.

[22] H. Wang, X. Bai, M. Yang, S. Zhu, J. Wang and W. Liu, "Scene Text Retrieval via Joint Text Detection and Similarity Learning", In Proc. CVPR, pp 4558-4567, 2021.

[23] Y. Wu, W. Liu and S. Wan, "Multiple attention encoded cascade R-CNN for scene text detection", Journal of Visual Communication and Image Representation, 103261, 2021.

[24] Z. Cheng, J. Liu, B. Zou, L. Qiao, Y. Xu, S. Pu, Y. Niu, F. Wu and S. Zhou, "FREE: A fast and robust end-to-end video text spotter", IEEE Trans. IP, pp 822-837, 2021.

[25] P. N. Chowdhury, P. Shivakumara, R. Raghavendra, S. Nag, U. Pal, T. Lu and D. Lopresti, "An episodic learning network for text detection on human bodies and sports images", IEEE Trans. CSVT, 2021.

[26] Y. Gao, X. Li, J. Zhang, Y. Zhou, D. Jin, J. Wang, S. Zhu and X. Bai, "Video text tracking with a spatio-temporal complementary model", IEEE Trans. IP, pp 9321-9331, 2021.

[27] L. Nandanwar, P. Shivakumara, R. Ramachandra, T. Lu, U. Pal. A. Antonacopoulos and Y. Lu, "A mew deep wavefront based model for text localization in 3D video", IEEE Trans. CSVT, 2021.

[28] C. V. Loc, J. C. Brurie and J. M. Ogier, "Document images watermarking for security issue using fully covolutional networks", In Proc. ICPR, pp 1091-1096, 2018.

[29] C. V. Loc, J. C. Brurie and J. M. Ogier, "Stable regions and object fill-based approach for document images watermarking", In Proc. DAS, pp 181-186, 2018.

[30] P. N. Chowdhury, P. Shivakumara, R. Raghavendra, U. Pal, T. Lu and M. Blumenstein, "A new U-net based license plate enhancement model in night and day images", In Proc. ACPR. pp 749-763, 2019.

[31] L. Tychsen-Smith and L. Petersson, "Improving object localization with fitness nms and bounded iou loss", In Proc. CVPR, pp. 6877-6885, 2018.

[32] J. Yang, Q. Liu and K. Zhang, "Stacked hourglass network for robust facial landmark localisation", In Proc. CVPR, pp 79-87, 2017.

[33] G. Ning, P. Liu, X. Fan and C. Zhang, "A top-down approach to articulated human pose estimation and tracking", In Proc. ECCV, 2018.

[34] X. Wang, K. C. Chan, K. Yu, C. Dong and C. C. Loy, "Edvr: Video restoration with enhanced deformable convolutional networks", In Proc. CVPRW, 2019.

[35] H. Xu, L. Yao, W. Zhang, X. Liang and Z. Li, "Auto-fpn: Automatic network architecture adaptation for object detection beyond classification", In Proc. ICCV, pp 6649-6658, 2019.

[36] Y. Zhou, Z. Zhang and D. Ding, "Partially-coupled nonlinear parameter optimization algorithm for a class of multivariate hybrid models", Applied Mathematics and Computation, *414*, p.126663, 2022.

[37] Z. Zhang, "Improved adam optimizer for deep neural networks", In Proc. IWQoS, 2018.

[38] N. Nayef, Y. Patel, M. Busta, P. N. Chowdhury, D. Karatzas, W. Khlif, … and J. M. Ogier, "ICDAR2019 robust reading challenge on multi-lingual scene text detection and recognition—RRC-MLT-2019", In Proc. ICDAR, pp 1582-1587, 2019.

[39] T. L. Yuan, Z. Zhu, K. Xu, C. J. Li, T. J. Mu and S. M. Hu, "A large chinese text dataset in the wild", Journal of Computer Science and Technology, pp 509-521, 2019.

[40] C. K. Ch'ng, and C. S. Chan, "Total-text: A comprehensive dataset for scene text detection and recognition", In Proc. ICDAR, pp 935-942, 2017.