DCT-DWT-FFT Based Method for Text Detection in Underwater Images

Ayan Banerjee[1], Palaiahnakote Shivakumara[2(✉)], Soumyajit Pal[1], Umapada Pal[1], and Cheng-Lin Liu[3,4]

[1] Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India
ab2141@cse.jgec.ac.in, umapada@isical.ac.in

[2] Department of Computer System and Technology, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia
shiva@um.edu.my

[3] National Laboratory of Pattern Recognition, Institute of Automation of Chinese Academy of Sciences, Beijing, China
liucl@nlpr.ia.ac.cn

[4] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

Abstract. Text detection in underwater images is an open challenge because of the distortions caused by refraction, absorption of light, particles, and variations depending on depth, color, and nature of water. Unlike existing methods aimed at text detection in natural scene images, in this paper, we have proposed a novel method for text detection in underwater images through a new enhancement model. Based on observations that fine details of text in image share with high energy, spatial resolution, and brightness, we consider Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT), and Fast Fourier Transform (FFT) for image enhancement to highlight the text features. The enhanced image is fed to a modified Character Region Awareness for Text Detection (CRAFT) model to detect text in underwater images. To explore enhancement methods, we evaluate six combinations of image enhancement techniques, namely, DCT-DWT-FFT, DCT-FFT-DWT, DWT-DCT-FFT, DWT-FFT-DCT, FFT-DCT-DWT, FFT-DWT-DCT. Experimental results on our dataset of underwater images and benchmark datasets of natural scene text detection, namely, MSRA-TD500, ICDAR 2019 MLT, ICDAR 2019 ArT, Total-Text, CTW1500, and COCO Text show that the proposed method performs well for both underwater and natural scene images and outperforms the existing methods on all the datasets.

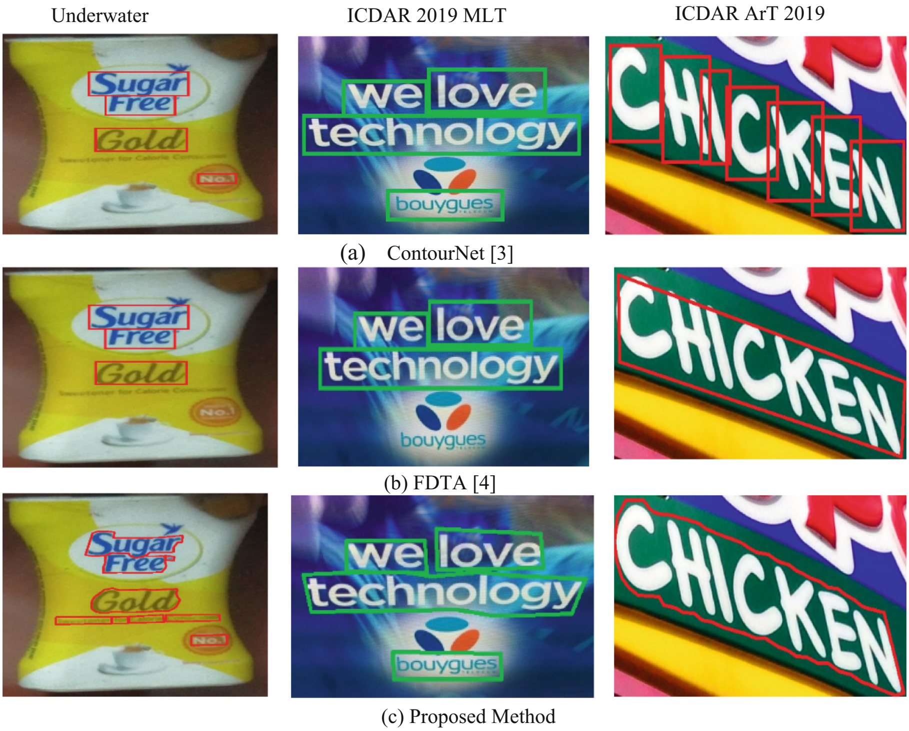
Keywords: Under water images · Text detection in underwater images · Image enhancement · Discrete cosine transform · Wavelet transform · Fourier transform · Modified Character Region Awareness for Text Detection (CRAFT)

1 Introduction

At present the tourism sector has been extended to perform activities, such as games, dining, marriage rituals, undersea, river, and water. Therefore, image indexing, retrieval, and

C. Wallraven et al. (Eds.): ACPR 2021, LNCS 13189, pp. 218–233, 2022.
https://doi.org/10.1007/978-3-031-02444-3_16

understanding of underwater images have received increasing attention from researchers in the field of image processing and computer vision [1, 2]. For example, in the case of scuba diving under the sea, text detection approach can be used to trace the swimmer by using text appeared on equipment and camera, such that we can prevent them to reaching dangerous area. At the same time, guide can control and monitor the swimmer and teach them different skills. In the similar way, text detection can be used to retrieve under water activities in the ocean. Therefore, text detection in underwater images is useful and significant in understanding underwater images and videos. Compared to text detection in natural scene images, detection in underwater images is more challenging due to various distortions caused by the refraction of light, the surface of water, depth of water, and particles in water. Many methods have been proposed for text detection in natural scene images [3, 4] but text detection in underwater images has received little attention. Existing natural scene text detection methods do not perform well on underwater images. In Fig. 1, the results of two state-of-the-art methods, ContourNet [3] and FDTA [4] are shown where it can be seen that tiny text lines in underwater images are missed despite their excellence performance on natural scene images. This is because underwater images lose quality due to distortion caused by the nature of water. On the other hand, the proposed method can detect such text properly in underwater as well as


Underwater ICDAR 2019 MLT ICDAR ArT 2019

(a) ContourNet [3]

(b) FDTA [4]

(c) Proposed Method

Fig. 1. Example of text detection of proposed and existing methods for under water and natural scene images. Results of ContourNet, FDTA and the proposed method are shown in (a), (b) and (c), respectively.

natural scene images. This is the contribution of the proposed model compared to the state-of-the-art methods.

In order to address the challenges of underwater images, inspired by the work [5], which combine different frequency domain analysis methods to improve the image quality for face anti-spoofing detection, we explore different combinations of Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT) and Fast Fourier Transform (FFT) to enhance the text details in underwater images. The intuition behind the idea is that the text pixels in any image share high energy, spatial resolution, and brightness compared to non-text pixels. It is noted that the energy can be achieved by DCT, the fine scaling can be achieved by DWT and the brightness is achieved by FFT. The combination of DCT-DWT-FFT produces six enhanced images. The enhanced images are supplied to modified Character Region Awareness for Text Detection (CRAFT) [6] in such a way that the model should work for both underwater and natural scene images.

In summary, the main contributions of the proposed method are as follows. (i) This is the first work addressing the challenges of text detection in underwater images. (ii) The combination of DCT-DWT-FFT is introduced for enhancing low contrast text information in underwater images. (iii) The proposed text detection method performs well on both underwater and natural scene images and outperforms state-of-the-art methods on multiple datasets.

2 Related Work

The existing methods on text detection in natural scene images can be categorized into top-down (regression based), bottom-up (segmentation based), and hybrid models for accurate text detection. We review the recent methods of the same.

The regression-based methods consider the whole text as an object for text detection. For instance, Cao et al. [4] proposed FDTA, which is fully convolutional scene text detection with text attention. Liu et al. [7] proposed GCCNet, which is a grouped channel composition network for scene text detection in natural scene images. The model optimizes anchor functions rather than a handcrafted feature for tackling the challenges of text detection. Shi et al. [8] explored iterative polynomial parameter regression for accurate arbitrarily shaped scene text detection. However, these methods are not robust for handling an arbitrarily shaped text.

To overcome the problem of arbitrarily-orientation and arbitrarily shaped text, the segmentation-based methods were developed. These methods used character and pixel information as local information for accurate text detection. For instance, Qin et al. [9] explored soft attention mechanism and dilated convolution for detecting arbitrarily shaped text in natural scene images. Baek et al. [6] developed a model called CRAFT, which is character region awareness for text detection in natural scene images. Dai et al. [10] proposed scale-aware data augmentation and shape similarity constant for accurate text detection in natural scene images. Hu et al. [11] proposed TATD, which is text contour attention for scene text detection in natural scene images. The model uses text center intensity maps and text kernel maps for accurate results. Liao et al. [12] proposed MaskTextSpotter, which is a trainable neural network for spotting text with arbitrary shapes. The model works based on developing sequence to sequence network. Deng et al. [13] developed a method called RFRN, which is recurrent features refinement network for accurate and efficient text detection in natural scene images. However, these methods are sensitive to distortion and complex background images.

To overcome the problems of regression and segmentation-based methods, hybrid methods were developed. These methods consider merits of regression and segmentation-based approaches to improve text detection performance. For example, Wang et al. [3] proposed ContourNet for detecting text in natural scene images that use advantages of both regressions-based models and segmentation-based models to tackle the challenge of arbitrarily shaped text detection. Liu et al. [14] proposed semi-supervised learning for text detection in natural scene images.

In summary, most of the models targeted the challenges of text detection in natural scene images but not other images, such as low light images, deformable text detection, tattoo text detection, and underwater images. However, there are methods [15–17] that consider the combination of enhancement and deep learning for low light images, deformable text detection from sports images based on episodic learning, and tattoo text detection based on deformable convolutional inception neural network. None of the existing methods consider underwater images for text detection. Hence, this work aims to develop an enhancement model for improving text detection in underwater and natural scene images.

3 Proposed Method

To detect text in underwater and natural scene images, it is observed that the properties of pixels, such as energy, fine scaling (spatial resolution), and brightness are common for all the text pixels regardless of image type and irrespective of qualities. Inspired by the method [5] where the combinations of DWT-LBP-FFT have been used for separating pixels affected by face attack from the actual pixels, we explore the combination in a different way for enhancing text pixels in the underwater images. In addition, it is noted that the proposed transforms involve combination of low and high pass filters to find fine details, namely edge information for reconstructing images. This observation motivated us to propose different combinations of above transforms such that the fine details in poor quality under water image caused by multiple adverse factors can be enhanced.

To explore the optimal configuration of image enhancement model, we evaluate different combinations of three image transform techniques: DCT, DWT, and FFT. The combinations of transforms results in six versions of enhanced images, namely, DCT-DWT-FFT, DCT-FFT-DWT, DWT-DCT-FFT, DWT-FFT-DCT, FFT-DCT-DWT, FFT-DWT-DCT. After image enhancement, we adopt the state-of-the-art text detection method, Character Region Awareness for Text Detection (CRAFT) [6], which works well for good quality images by studying character shape. We use the same to modify the model such that the modified model can withstand the challenges of text detection in underwater images by considering six enhanced images as input in this work. The schematic diagram of the proposed work is shown in Fig. 2.

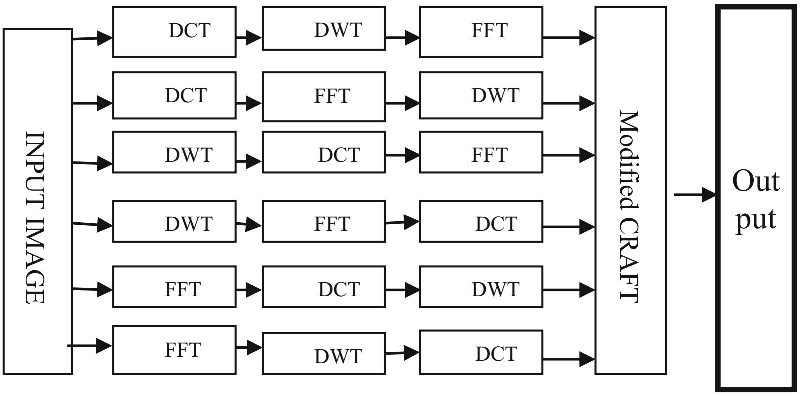


Fig. 2. The proposed block diagram for text detection in underwater images.

3.1 DCT-DFT-FFT Images for Enhancement

For the input image, the proposed method obtains IDCT, IDWT, and IFFT images. In order to take advantage of DCT, DWT, and FFT, the proposed approach performs the combination of DCT, DWT, and FFT bypassing reconstructed images. For example, the reconstructed image of DCT is supplied to DWT, the reconstructed image of DWT is supplied to FFT, which outputs the final reconstructed image of the first combination. The same process is continued to obtain all the six reconstructed images by the six combinations, namely, DCT-DWT-FFT, DCT-FFT-DWT, DWT-DCT-FFT, DWT-FFT-DCT, FFT-DCT-DWT, FFT-DWT-DCT. The calculations of DCT, DWT and FFT are obtained as defined in Eqs. (1)–(3), respectively.

$$G(r,s) = \beta_r\beta_s \sum\nolimits_{x=0}^{P-1}\sum\nolimits_{y=0}^{Q-1} g(x,y)cos\left[\frac{\pi(2x+1)r}{2P}\right]cos\left[\frac{\pi(2y+1)s}{2Q}\right], \tag{1}$$

$$\begin{aligned} g(x,y) = {} & \frac{1}{\sqrt{PQ}}\sum_{u}\sum_{v} W_\delta(k_0,u,v)\delta_{k_0,u,v}(x,y) \\ & + \frac{1}{\sqrt{PQ}}\sum_{i=H,V,D}\sum_{k=k_0}\sum_{u}\sum_{v} W_\omega^i(k,u,v)\omega_{k,u,v}^i(x,y), \end{aligned} \tag{2}$$

$$B_j = \sum_{u=0}^{U-1} e^{-i\frac{2\pi ju}{U}} b_u, \tag{3}$$

where, $\beta_r and \beta_s$ represent the beta distribution of the pixel, P and Q are the spectrum intensity and spectrum density, respectively. On the other hand, W_δ is the wavelet transform function k, u, v represents the horizontal, vertical and diagonal transform, respectively. Last but not the least, b_u depicts the pixel value before DFT. Fast Fourier Transform (FFT) is a fast way of computing Discrete Fourier Transform by taking 2-point and 4-point DFT and generalizing them to 8-point, 16-point, …, 2^{r}-point.

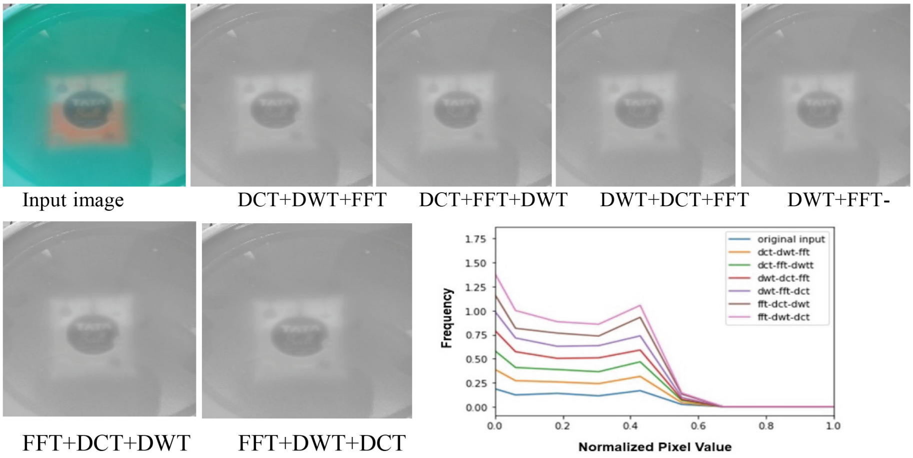

Input image
DCT+DWT+FFT
DCT+FFT+DWT
DWT+DCT+FFT
DWT+FFT-
FFT+DCT+DWT
FFT+DWT+DCT

Fig. 3. Six combinations of enhanced images and their pixel distributions.

The results of each combination can be seen in Fig. 3 for the input image, where it can be seen that the brightness increase for all the enhanced images. We believe that each combination helps us to enhance the fine details of text pixels in the underwater images. As a result, the contrast between text and non-text pixel increases. It is evident from the plots of six enhanced images shown in Fig. 3. Here, X-axis represents the normalized pixel value and Y-axis represents the pixel frequency downscaled by 100. Higher pixel frequency means high degree of enhancement which helps for better detection. It can be seen that the pixel frequency of six enhanced images increases compared to the values in the input image. This observation motivated us to explore CRAFT mode for text detection from the enhanced images. Therefore, we modify the CRAFT such that it accepts all the six enhanced images as input for accurate text detection irrespective of image type and quality.

3.2 Proposed Modified CRAFT for Text Detection in Underwater Images

It is noted that the existing CRAFT can address most of the challenges, such as arbitrarily shaped, arbitrarily-orientation, which are common in the case of underwater images. However, it works well for the images with good quality and contrast but not for the underwater image which generally suffers from poor quality. Therefore, we modify the CRAFT such that it performs well for underwater images by considering all the six enhanced images obtained from the previous section as input. The modified architecture can be seen in Fig. 4. The ResNet-50 has been used here as the foundation of CRAN. We use FPN to meld include maps produced by various phases of the backbone initializing from the top. Utilizing the melded highlight maps, the consideration module further upgrades its discriminative parts by creating comparing consideration loads. Then the improved element maps are used in the correction module and the acknowledgment organization to create the character groups. The proposed network is an end-to-end trainable and the acknowledgment network utilized is equivalent to the consideration-based encoder-decoder in [18].

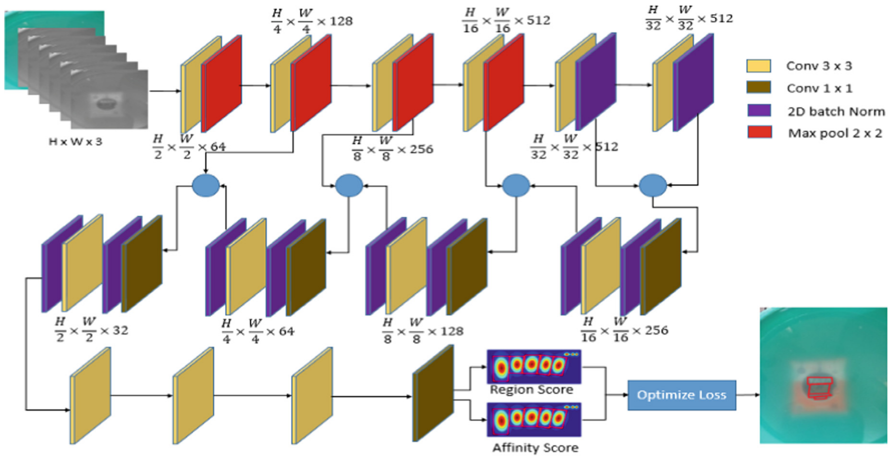

$\frac{H}{4} \times \frac{W}{4} \times 128$
$\frac{H}{16} \times \frac{W}{16} \times 512$
$\frac{H}{32} \times \frac{W}{32} \times 512$
🟨 Conv 3 x 3
🟫 Conv 1 x 1
🟪 2D batch Norm
🟥 Max pool 2 x 2
H x W x 3
$\frac{H}{2} \times \frac{W}{2} \times 64$
$\frac{H}{8} \times \frac{W}{8} \times 256$
$\frac{H}{32} \times \frac{W}{32} \times 512$
$\frac{H}{2} \times \frac{W}{2} \times 32$
$\frac{H}{4} \times \frac{W}{4} \times 64$
$\frac{H}{8} \times \frac{W}{8} \times 128$
$\frac{H}{16} \times \frac{W}{16} \times 256$
Region Score
Affinity Score
Optimize Loss

Fig. 4. Proposed modified CRAFT for text detection in under water image.

The 2D character detection module comprises a detection head, a component savvy duplication (i.e. element-wise multiplication) activity. Given the information highlight maps I_{in}, the consideration head produces the consideration map, CM through three convolution layers, every one of which is trailed by a cluster standardization (i.e. batch normalization) layer and a ReLU activation layer. Then, at that point the yield highlight maps I_{out} can be registered as defined in Eq. (4).

$$I_{out} = I_{in} \otimes CM \tag{4}$$

where $\otimes$ addresses component insightful augmentation. Note that CM has one channel and therefore, the proposed model broadcast it to a similar shape as I_{in} to accomplish component astute augmentation. Even though our consideration module is exceptionally basic, it brings extensive improvement because of the oversight of the consideration map during preparation.

We use the correction module to amend the element guides of discretionary formed examples to customary ones. It comprises an amendment head organization and a sampler. In the first place, the correction head predicts the directions of K control focuses, then, at that point the sampler utilizes Thin-Plate-Spline (TPS) change to produce the redressed include maps. Unique concerning Aster that straightforwardly redresses the info pictures, we amend the upgraded highlight maps. As enough discriminative data is separated, we don't require as many convolution layers as Aster. The configuration of our modified CRAFT architecture is shown in Table 1.

Table 1. Configuration of the proposed modified CRAFT architecture

Layer name	Configuration	Out size
Conv 1	7 × 7; 2 × 2; 2 × 2; 64, maxpool: 2 × 2; 2 × 2; 0 × 0	8 × 16
Conv2	5 × 5; 1 × 1; 1 × 1; 128, maxpool: 2 × 2; 2 × 2; 0 × 0	4 × 8
Conv3	3 × 3; 1 × 1; ×1; 256, maxpool: 2 × 2; 2 × 2; 0 × 0	2 × 4
Conv4	1 × 1; 1 × 1; ×1; 512, maxpool: 2 × 2; 2 × 2; 0 × 0	1 × 4
fc1	512	512
fc2	$K \times 3$	$K \times 3$

During preparation, the consideration module and acknowledgment module are regulated. The entire framework can be prepared in a start to finish way, with the accompanying loss function as defined in Eq. (5).

$$L = \sum\nolimits_{I \in textImage} \lambda \times L_{att} + L_{rec}, \tag{5}$$

where, L_{att} addresses the deviation between the anticipated consideration map and the ground truth, which is determined by the Smooth L_1 Loss as defined in Eq. (6).

$$\begin{aligned} Smooth_{L_1}(x) &= \{-0.5a^2 if\ a < 1; \\ &= \{|\mathrm{a}| - 0.5 \text{ otherwise} \end{aligned} \tag{6}$$

So, L_{att} can be defined using Eq. (7).

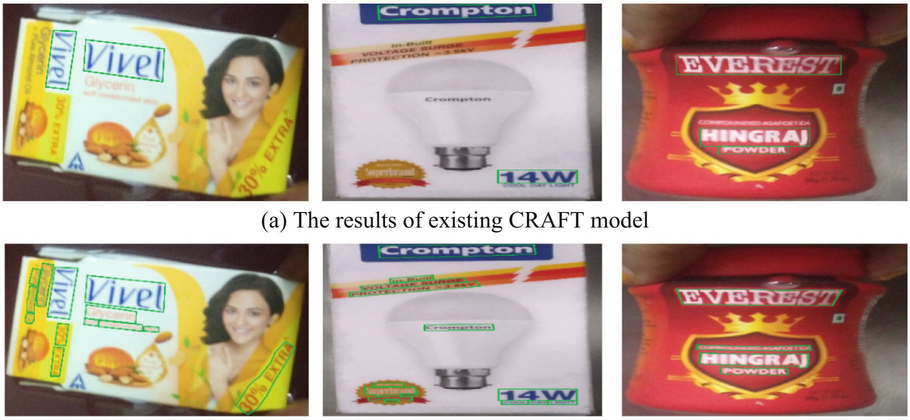
$$L_{att} = Smooth_{L_1}\left(CM - CM^*\right), \tag{7}$$

where CM is the anticipated consideration guide and CM^* is the comparing ground truth. Moreover, L_{rec} denotes the recognition loss, which can be formulated as defined in Eq. (8).

$$L_{rec} = -\frac{1}{N}\sum\nolimits_{i=1}^{N} \log p(x_i|I, \emptyset), \tag{8}$$

where $x_1, x_2, \ldots\ldots\ldots, x_N$ is the ground truth record grouping, ∅ addresses all teachable boundaries of our organization, I is the info picture. The hyper-boundary λ is intended to balance two losses. Samples are chosen from SynthText and Synth90K is utilized during preparation. For tests from SynthText, we use jumping box comments of each character to create the ground reality of the consideration map. Since Synth90K doesn't provide character-level comments, we disregard L_{att} for tests from Synth90k, i.e., the acknowledgment misfortune is utilized to upgrade the model. λ is observationally set to 1000 in our tests.

The results of the proposed modified CRAFT and the existing CRAFT are shown in Fig. 5(a)–(b), where it can be seen that the proposed modified CRAFT detects all the text in all the three images including tiny and a big text as shown in Fig. 5(b), while the existing CRAFT misses' text in underwater images especially for tiny text as shown in Fig. 5(a). This shows that modifications to existing CRAFT are effective for accurate text detection in underwater images.



(a) The results of existing CRAFT model

(b) The results of the proposed modified CRAFT model

Fig. 5. Effectiveness of the proposed modified CRAFT for different water images.

4 Experimental Results

Creating a dataset for text detection in underwater images is not an easy task and at present there is no dataset for underwater images. Therefore, we created our own dataset by immersing objects containing text in the water at different levels. Both clear and polluted water are used for data collection. For making the dataset complex, we have added dust and mud into water. In addition, we used different objects, papers, bottles with labels, different covers of the packets for creating the dataset. The text in such images can have arbitrary-shaped characters, orientation, and dense text in addition to the adverse effect of water. We believe that the way we created the dataset matches with the real scenario of underwater images. Our dataset consists of 500 images with large variations.

To show that the proposed model works well for natural scene images, we consider six benchmark datasets as follows. **MSRA-TD500** [19]**:** This is to test the ability of multi-oriented and multi-lingual text detection. The dataset provides 300 images for training and 200 images for testing. **CTW1500** [19]: This is to test the curved text detection ability of the methods. It provides 1000 images for training and 500 images for testing. **Total-Text** [19]: This is also a curved text dataset with more variations in the images for evaluating the performance of the methods. It provides 1255 images for training and 300 for testing. **ICDAR 2017 MLT** [19]: This is to test the multi-lingual ability of the methods, which includes 9 different languages. It provides 7200 images for training images, 1800 images for validation, and 9000 images for testing. **ICDAR 2019 ArT dataset** [20]**:** This dataset combines the images of Total-Text, CTW1500, and Baidu Curved Scene Text. This is a huge dataset compared to all other datasets. In total, 10,166 images are there, and 5603 images are used for training and 4563 images for testing. **COCO-Text** [19]: This is not created to evaluate text detection methods; the images are collected with the intention of other objectives. As a result, one can expect large variations in the images compared to other benchmark datasets. It provides 43686 images for training, 20000 images for validation, and 900 images for testing.

To show the superiority of the proposed method over existing methods, we compared the results of the proposed method with the results of the SOTA methods [3, 4, 6, 8, 9, 14]. The reason to choose the above existing methods for comparison is that the objective of these methods is the same as the proposed work. In addition, the methods addressed the challenges that similar to text detection in underwater images. To test the above-mentioned existing methods on our underwater image dataset, we retrain the methods with training samples of respective datasets. The standard measures, namely, Precision (P), Recall (R), and F-Score (F) for evaluating the performance of the methods are used as defined in [3–8].

For experiments, we use 70:30 ratios for training and testing in the case of our underwater image dataset while for all other benchmark datasets, we use the number of training and testing samples according to the ratio provided in the respective datasets. However, the evaluation scheme followed in this study is the same for all the experiments.

4.1 Ablation Study

The key steps of the proposed method are the combination of DCT-DWT-FFT and modifications to the existing CRAFT to achieve the best text detection performance. To validate the contribution of each transform, enhanced images obtained by six combinations, and the effectiveness of the proposed modified CRAFT, the following experiments are conducted.

(i) The existing CRAFT without any modifications was applied on an underwater image dataset for calculating measures as reported in Table 2 and the results are considered as baseline results for comparing with other steps of the proposed method. In this experiment, the input underwater images are passed to the existing CRAFT for text detection. (ii)–(iv) The reconstructed images given by DCT, DWT, and FFT are fed to the proposed modified CRAFT for text detection. This is to test the contribution of DCT alone, DFT alone and FFT alone to achieve the best detection results by the proposed method. (v)–(x) Enhanced images given by each combination are supplied to proposed modified CRAFT for text detection. This is to test the effectiveness of each combination. (xi) The input images are passed to the proposed modified CRAFT without enhancement images. This is to test the contribution of the modifications done to the existing CRAFT. (xii) All the six enhanced images are fed to the proposed modified CRAFT for text detection in underwater images.

It is observed from Table 2 that the results of experiments from (ii)–(xi) show the Precision, Recall and F-Score of each experiment are improving compared to the baseline results of the experiment (i). Therefore, one can infer that the transforms, the combinations of different transforms, and modifications to the existing CRAFT are all effective and contribute equally to achieving the best results for text detection in underwater images by the proposed method as reported in Experiment (xii).

Table 2. Ablation Study using our underwater images dataset.

#	Experiments	P	R	F
(i)	Baseline method: Existing CRAFT [6]	60.3	61.8	60.1
(ii)	DCT+ Proposed modified CRAFT	62.7	62.2	62.3
(iii)	DWT+ Proposed modified CRAFT	64.3	65.8	65.1
(iv)	FFT+ Proposed modified CRAFT	66.7	66.2	66.3
(v)	DCT-FFT-DWT+ Proposed modified CRAFT	70.3	71.8	71.1
(vi)	DWT-DCT-FFT+ Proposed modified CRAFT	72.7	72.2	72.3
(vii)	DWT-FFT-DCT+ Proposed modified CRAFT	76.3	75.8	76.1
(viii)	FFT-DCT-DWT+ Proposed modified CRAFT	82.7	82.2	82.3
(ix)	FFT-DWT-DCT+ Proposed modified CRAFT	84.4	84.9	84.2
(x)	DCT-DWT-FFT+ Proposed modified CRAFT	86.8	86.3	86.4
(xi)	Proposed modified CRAFT (without enhancement)	87.2	87.1	87.1
(xii)	Proposed method	**89.2**	**88.7**	**89.1**

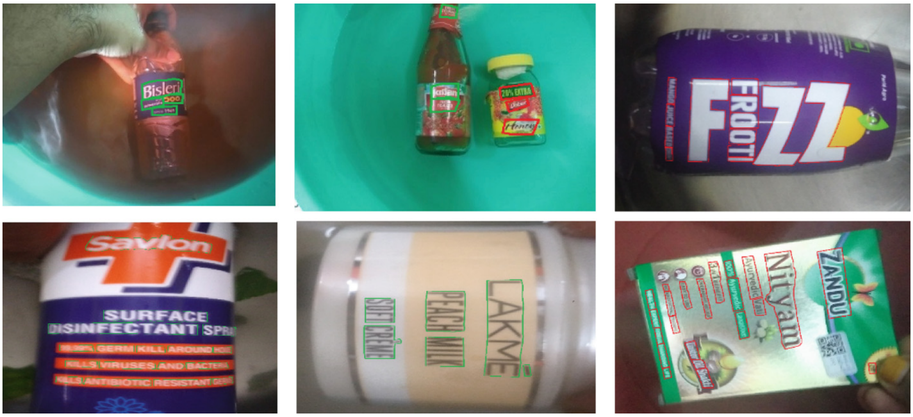


Fig. 6. Text detection results of the proposed method for underwater images.

4.2 Experiments on Our Underwater Images Dataset

Sample results of the proposed method for text detection in underwater images are shown in Fig. 6, where it can be seen that the proposed method is capable of detecting the different types of texts in the underwater images including tiny, dense, arbitrarily oriented text with complex background. Therefore, one can argue that the proposed model is robust to underwater images of different qualities. For quantitative results, to show that the enhancement step presented in Sect. 3.1 is effective in improving the text detection performance of the methods, we calculate the measures by feeding input image as input for the proposed and existing methods, which is called before enhancement experiments. Similarly, the measures are calculated by feeding six enhanced obtained by enhancement steps as input to the proposed and existing methods, which is called after enhancement experiments.

In the case of before-enhancement experiments, the input images are fed to the proposed modified CRAFT without enhanced images for text detection. For after-enhancement experiments, the six enhanced images are fed to the proposed modified CRAFT for text detection. It is observed from Table 3 that all the methods report better results in terms of Precision, Recall, and F-Score for after-enhancement compared to before-enhancement. This indicates that the enhancement step is effective and contributes to achieving better detection results for underwater images. In the same way, when we compare the results of the proposed and existing methods before and after enhancement, the proposed method is the best at Precision, Recall and F-Score compared to the existing methods. Therefore, one can conclude that the proposed method is capable of addressing the challenges of underwater images. On the other hand, since the existing methods were developed for detecting text in natural scene images, the existing methods are not effective for underwater images, which are affected by distortion caused by the depth of water, purity of water, light refraction, light absorption, and the objects like labels on the bottles, covers, papers, different objects etc.

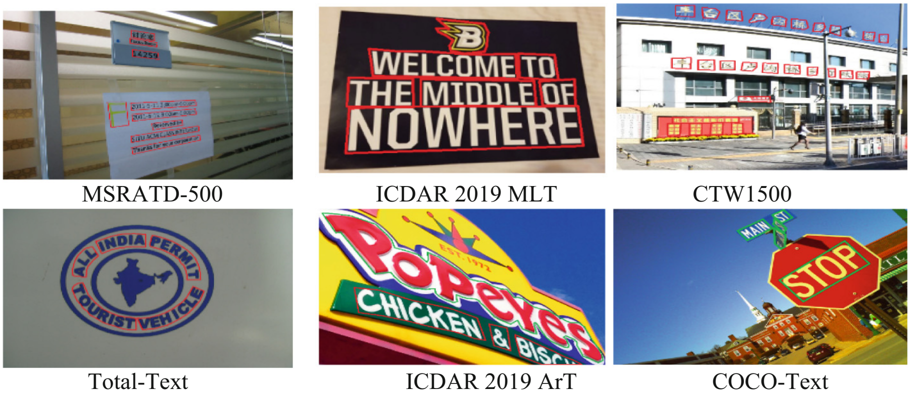
Table 3. Performance of the proposed and existing methods on underwater images dataset

Methods	Before enhancement			After enhancement		
	P	R	F	P	R	F
ContourNet [3]	69.7	65.9	67.8	71.7	67.9	69.8
FDTA [4]	77.9	80.1	79.4	79.9	82.1	81.4
Qin et al. [9]	84.7	83.2	84.2	85.7	84.2	85.2
Shi et al.[8]	83.9	84.3	84.1	84.9	85.3	85.1
SemiText [14]	84.2	85.5	84.7	85.2	86.5	85.7
CRAFT [6]	60.3	61.8	60.1	64.3	65.8	64.1
Proposed	**89.2**	**88.7**	**89.1**	**91.2**	**90.7**	**91.1**

4.3 Experiments on Benchmark Dataset of Natural Scene Images

To show that the proposed method has ability to detect text in natural scene images, the measures are calculated for the images of six standard natural scene text datasets, namely, MSRA-TD500, ICDAR 2019 MLT, CTW1500, Total-Text, ICDAR 2019 ArT, and COCO-Text. Sample results of the proposed method are shown in Fig. 7, where we can observe that the proposed method detects text well for all the images of different datasets. This shows that despite the proposed model is developed for text detection in underwater images, it detects text well for natural scene images and hence the proposed method is robust.

Quantitative results of the proposed and existing methods before, after-enhancement for all the aforementioned datasets are reported in Tables 4 and 5. Tables 4 and 5 show that all the methods including the proposed method report high results for after-enhancement (After) compared to before-enhancement (Before) in terms of all the three measures. Therefore, we can confirm that the proposed enhancement is useful for improving text detection performance even for natural scene images also. Similarly, the results of the proposed and existing methods after-enhancement show that the proposed method is better than existing methods. This indicates that the proposed method is independent of image type, text type, and image quality. It is evident from the results of the proposed method on all the datasets that the results are almost the same for all the datasets. This is the advantage of obtaining the enhanced images by the six combinations of DCT-DWT-FFT and modified CRAFT. The reason for the poor results of existing methods is that although the models are robust to low contrast, low resolution, and taking advantage of deep learning, the models are not consistent and stable when the images suffer from poor quality affected by multiple adverse factors of water images.Overall, the proposed enhancement is effective in improving the performance of text detection for both underwater and natural scene images. In addition, the proposed model is generic because it performs well for images of different complexities.



MSRATD-500 ICDAR 2019 MLT CTW1500

Total-Text ICDAR 2019 ArT COCO-Text

Fig. 7. Example of text detection of the proposed model for images of different benchmark natural scene text datasets.

Sometimes, when the text is too tiny and water contains more dust as shown in Fig. 8, the proposed model does not detect text, accurately. It can be seen from the examples shown in Fig. 8, where the proposed model misses text and does not fix proper bounding boxes for each text line in the images. Therefore, there is a scope for improving the proposed model further. In these cases, just enhancement using image information is not sufficient. We need to extract object information to define the context and then the context information can be used to enhance the whole region rather than focusing only on text information. Next, to improve the quality of the enhanced region, one can think of super-resolution concept, which enhances the fine details of the text.

Table 4. Text detection performance of the proposed and existing methods on MSRATD-500, ICDAR 2019 MLT and CTW1500.

Methods	MSRA-TD500						ICDAR 2019 MLT						CTW1500					
	Before			After			Before			After			Before			After		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
ContourNet [3]	68.7	68.9	68.8	78.7	78.9	78.8	79.7	75.9	77.8	81.7	77.9	79.8	84.6	82.7	83.7	86.6	84.7	85.7
FDTA [4]	77.9	86.1	83.4	79.9	88.1	85.4	77.1	80.2	79.3	79.1	82.2	81.3	83.1	79.9	81.5	85.1	81.9	83.5
Qin et al. [9]	86.7	84.2	85.2	88.7	86.2	87.2	84.4	83.5	84.6	86.4	85.5	85.6	84.9	79.6	82.3	86.9	81.6	84.3
Shi et al.[8]	85.9	85.3	85.1	87.9	87.3	87.1	83.7	84.8	84.9	85.7	86.8	85.9	86.1	81.2	83.4	88.1	83.2	85.4
SemiText [14]	84.2	86.5	85.7	86.2	88.5	87.7	84.0	85.1	84.2	86.0	87.1	86.2	85.8	83.1	84.5	87.8	85.1	86.5
CRAFT [6]	80.3	81.8	80.1	82.3	83.8	82.1	80.8	81.8	81.2	82.8	83.8	83.2	80.8	81.8	81.2	82.8	83.8	83.2
Proposed	**89.3**	**88.8**	**89.2**	**91.3**	**90.8**	**91.2**	**89.3**	**89.4**	**89.5**	**91.3**	**91.4**	**91.5**	**89.6**	**89.1**	**89.3**	**91.6**	**91.1**	**91.3**

Table 5. Text detection performance of the proposed and existing methods on Total-Text, ICDAR 2019 ArT and COCO-text datasets.

Methods	Total-text						ICDAR 2019 ArT						COCO-text					
	Before			After			Before			After			Before			After		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
ContourNet [3]	82.6	74.6	78.5	84.6	76.7	80.6	84.1	82.7	83.3	86.1	84.7	85.3	87.7	85.9	86.8	89.7	87.9	88.8
FDTA [4]	85.7	75.8	80.4	87.7	77.8	82.4	83.2	79.8	81.4	84.2	81.8	83.4	79.9	82.1	81.4	81.9	84.1	83.4
Qin et al. [9]	81.3	79.8	80.7	83.3	81.8	82.7	84.3	79.9	82.5	86.3	81.9	84.5	86.7	84.2	85.2	88.7	86.2	87.2
Shi et al.[8]	84.1	77.8	80.9	86.1	79.8	82.9	86.4	81.0	83.6	88.4	83.0	85.6	83.8	84.7	84.4	85.8	86.7	86.4
SemiText [14]	87.5	79.8	83.7	89.5	81.8	85.7	85.5	83.1	84.7	87.5	85.1	86.7	84.5	85.3	84.2	86.5	87.3	86.2
CRAFT [6]	81.8	82.8	82.2	83.8	84.8	84.2	80.3	81.8	80.1	82.3	83.8	82.1	81.8	82.8	82.2	83.8	84.8	84.2
Proposed	**89.6**	**88.8**	**88.6**	**91.6**	**90.8**	**90.6**	**88.6**	**88.2**	**87.8**	**90.6**	**90.2**	**89.8**	**89.9**	**89.7**	**89.6**	**91.9**	**91.7**	**91.6**

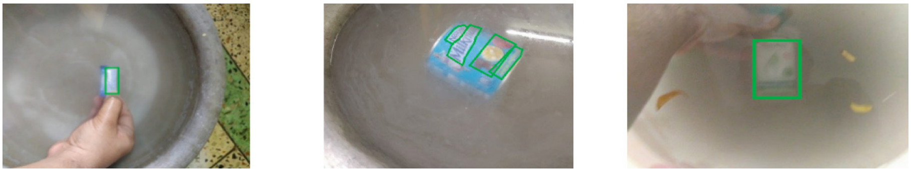


Fig. 8. Limitation of the proposed model.

5 Conclusion

We have proposed a novel method for text detection in underwater images through a new enhancement approach and using modifications of existing CRAFT. The main objective of the proposed work is to address the challenges of text detection in underwater images and natural scene images. To the best of our knowledge, this is the first work of its kind, unlike existing methods that focus only on text detection in natural scene images. For the enhancement, the proposed approach explores the combination of DCT-DWT-FFT, which generates six enhanced images for each input image. For text detection results, we have modified the existing CRAFT to detect text in underwater images irrespective of image type, text type, and quality affected by multiple adverse factors, which consider nine enhanced images as input. Experimental results of the proposed and existing methods on the underwater image dataset and six standard natural scene text datasets show that the proposed model is superior to existing methods in terms of consistency, stable results, and robustness to different datasets.

Acknowledgements. The work of Cheng-Lin Liu was supported by the National Key Research and Development Program under Grant No. 2018AAA0100400 and the National Natural Science Foundation of China under Grant No. 61721004. This work also partially supported by TIH, ISI, Kolkata.

References

1. Xue, M., et al.: Deep invariant texture features for water image classification. SN Appl. Sci. **2**(12), 1–19 (2020). https://doi.org/10.1007/s42452-020-03882-w
2. Kezebou, L., Oludare, V., Panetta, K., Againa, S.S.: Underwater object tracking benchmark and dataset. In: Proceedings of the HST (2019). https://doi.org/10.1109/HST47167.2019.9032954
3. Wang, Y., Xie, H., Zha, Z.J., Xing, M., Fu, Z., Zhang, Y.: ContourNet: Taking a further step toward accurate arbitrary-shaped scene text detection. In: Proceedings of the CVPR, pp. 11753–11762 (2020)
4. Cao, Y., Ma, S., Pan, H.: FDTA: Fully convolutional scene text detection with text attention. IEEE Access 155441–155449 (2020)
5. Zhang, W., Xiang, S.: Face anti-spoofing detection based on DWT-LBP-DCT features. Signal Process. Image Commun. (2020). https://doi.org/10.1016/j.image.2020.115990
6. Baek, Y., Lee, B., Han, D., Yun, S., Lee, H.: Character region awareness for text detection. In: Proceedings of the CVPR, pp. 9365–9374 (2019)
7. Liu, C., Yang, C., Hou, J.B., Wu, L.H., Zhu, X.B., Xiao, L.: GCCNet: Grouped channel composition network for scene text detection. Neurocomputing **454**, 135–151 (2021)
8. Shi, J., Chen, L., Su, F.: Accurate arbitrary-shaped scene text detection via iterative polynomial parameter regression. In: Ishikawa, H., Liu, C.-L., Pajdla, T., Shi, J. (eds.) ACCV 2020. LNCS, vol. 12624, pp. 241–256. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-69535-4_15
9. Qin, X., Jiang, J., Yuan, C.A., Qiao, S., Fan, W.: Arbitrary shape natural scene text detection method based on soft attention mechanism and dilated convolution. IEEE Access 122685–122694 (2020)
10. Dai, P., Li, Y., Zhang, H., Li, J., Cao, X.: Accurate scene text detection via scale-aware data augmentation and shape similarity constraint. IEEE Trans. Multim. (2021). https://doi.org/10.1109/TMM.2021.3073575
11. Hu, Z., Wu, X., Wang, J.: TCATD: text contour attention for scene text detection. In: Proceedings of the ICPR, pp. 1083–1088 (2021)
12. Liao, M., Lyu, P., He, M., Yao, C., Wu, W., Bai, X.: Mask TextSpotter: an end-to-end trainable neural network for spotting text with arbitrary shapes. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. LNCS, vol. 11218, pp. 71–88. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01264-9_5

13. Deng, G., Ming, Y., Xue, J.-H.: RFRN: A recurrent feature refinement network for accurate and efficient scene text detection. Neurocomputing **453**, 465–481 (2021)
14. Liu, J., Zhong, Q., Yuan, Y., Su, H., Du, B.: SemiText: scene text detection with semi-supervised learning. Neurocomputing **407**, 343–353 (2020)
15. Xue, M., et al.: Arbitrarily-oriented text detection in low light natural scene images. IEEE Trans. Multim. **23**, 2706–2720 (2020)
16. Chowdhury, P.N., et al.: A new episodic learning-based network for text detection on human body in sports images. IEEE Trans Circuits Syst. Video Technol. (2021). https://doi.org/10.1109/TCSVT.2021.3092713
17. Chowdhury, T., Shivakumara, P., Pal, U., Tong, L., Raghavendra, R., Chanda, S.: DCINN: deformable convolution and inception based neural network for tattoo text detection through skin region. In: Lladós, J., Lopresti, D., Uchida, S. (eds.) Document Analysis and Recognition – ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II, pp. 335–350. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-86331-9_22
18. Zhou, X., et al.: East: an efficient and accurate scene text detector. In: Proceedings of the CVPR, pp. 2642–2651 (2017)
19. Roy, S., Shivakumara, P., Pal, U., Lu, T., Kumar, G.H.: Delaunay triangulation-based text detection from multi-view images of natural scene. Pattern Recogn. Lett. **129**, 92–100 (2020)
20. Chng, C.K., Liu, Y., Sun, Y., Ng, C.C., Luo, C., Ni, Z.: ICDAR2019 robust reading challenge on arbitrarily-shaped text-RRC-ArT. In: Proceedings of the ICDAR, pp. 1571–1576 (2019)