Arbitrarily-Oriented Text Detection in Low Light Natural Scene Images

Minglong Xue, Palaiahnakote Shivakumara, Chao Zhang, Yao Xiao, Tong Lu, Umapada Pal, Daniel Lopresti, and Zhibo Yang

***Abstract*—Text detection in low light natural scene images is challenging due to poor image quality and low contrast. Unlike most existing methods that focus on well-lit (normally daylight) images, the proposed method considers much darker natural scene images. For this task, our method first integrates spatial and frequency domain features through fusion to enhance fine details in the image. Next, we use Maximally Stable Extremal Regions (MSER) for detecting text candidates from the enhanced images. We then introduce Cloud of Line Distribution (COLD) features, which capture the distribution of pixels of text candidates in the polar domain. The extracted features are sent to a Convolution Neural Network (CNN) to correct the bounding boxes for arbitrarily oriented text lines by removing false positives. Experiments are conducted on a dataset of low light images to evaluate the proposed enhancement step. The results show our approach is more effective compared to existing methods in terms of standard quality measures, namely, BRISQE, NIQE and PIQE. In addition, experimental results on a variety of standard benchmark datasets, namely, ICDAR 2013, ICDAR 2015, SVT, Total-Text, ICDAR 2017-MLT and CTW1500, show that the proposed approach not only produces better results for low light images, at the same time it is also competitive for daylight images.**

***Index Terms*—Image enhancement, gaussian pyramid filter, Homomorphic filter, COLD features, convolutional neural network, arbitrarily-oriented text detection.**

Manuscript received December 25, 2019; revised May 3, 2020 and July 7, 2020; accepted August 1, 2020. Date of publication August 7, 2020; date of current version August 24, 2021. The work of Tong Lu was supported by the Natural Science Foundation of China under Grants 61672273 and 61832008, in part by the Alibaba Group through Alibaba Innovative Research Program. The work of Palaiahnakote Shivakumara was supported by the Faculty Grant: GPF014D-2019, University of Malaya, Malaysia. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jianfei Cai. *(Corresponding author: Tong Lu.)*

Minglong Xue, Chao Zhang, Yao Xiao, and Tong Lu are with the National Key Lab for Novel Software Technology, Nanjing University, Nanjing 210000, China (e-mail: xueml@smail.nju.edu.cn; zhangchao_nju@126.com; iam_xiaoyao@126.com; lutong@nju.edu.cn).

Palaiahnakote Shivakumara is with the Department of Computer System and Information Technology, University of Malaya, Kuala Lumpur 43200, Malaysia (e-mail: shiva@um.edu.my).

Umapada Pal is with the Indian Statistical Institute, Kolkata 700001, India (e-mail: umapada@isical.ac.in).

Daniel Lopresti is with the Computer Science & Engineering, Lehigh University, Bethlehem, PA 18015 USA (e-mail: lopresti@cse.lehigh.edu).

Zhibo Yang is with the Alibaba Grroup, HangZhou 310000, China (e-mail: hibo.yzb@alibaba-inc.com).

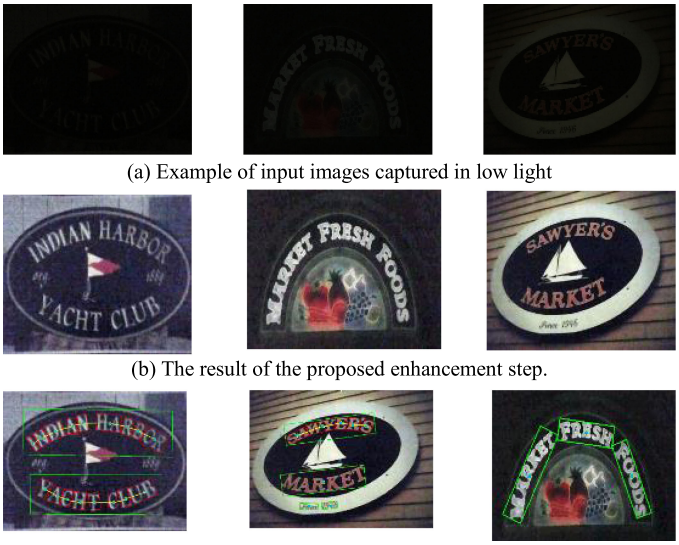
Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2020.3015037

I. Introduction

THERE are a number of situations that require capturing natural scene images in dark and low light environments, which include extreme weather such as fog, snow, and cloudy conditions [1], [2]. In these situations, captured images exhibit loss of quality, contrast and visibility, which result in poor performance for existing text detection methods. On top of this, arbitrarily text orientations which are common in such images make the problem even more complex and challenging. This motivates the work we present in this paper. In the past, powerful methods have been developed for addressing specific challenges such as images with complex backgrounds, low contrast, multiple scripts and orientations, curved text, different fonts and font sizes, etc. For example, Tian *et al.* [3] proposed scene text detection under weak supervision and addressed the challenge of multiple orientations, Deng *et al.* [4] proposed detecting scene texts via instance segmentation to address the problem of text and non-text pixel separation, Zhu *et al.* [5] proposed an efficient and accurate scene text detector and focused on addressing arbitrary orientations, Shi *et al.* [6] proposed a method to detect text with multiple orientations in natural images by linking segments, while Liu *et al.* [7] focused on curved text detection in the wild. Additionally, He *et al.* [8] proposed multi-oriented and multi-lingual scene text detection with direct regression. Xu *et al.* [9] proposed learning a deep direction field for irregular scene text detection. Raghunandan *et al.* [10] presented a method for text detection and recognition in born-digital video scene images. Tang *et al.* [11] described an approach for dense and arbitrarily-shaped scene text detection by instance-aware component grouping. Liao *et al.* [12] used a D ifferential Binarization Network (DBNet) for scene text detection in a real-time environment. Liu *et al.* [13] proposed an approach for scene text spotting through the use of an adaptive Bezier curve network. Zhan *et al.* [14] explored a Geometry Aware Domain Adaptation Network (GA-DAN) for scene text detection and recognition.

It can be noted from the above that past methods have used deep learning models for finding solutions to complex issues in order to improve text detection performance. However, none of these methods have considered the challenges presented by images captured in low light conditions, as illustrated in Fig. 1(a). As a result, it their performance is likely to suffer. To deal with such cases, it is helpful to enhance edges that preserve the shapes of characters and text lines as shown in Fig. 1(b), thereby making the content more visible. Working off of this enhancement,



(a) Example of input images captured in low light

(b) The result of the proposed enhancement step.

(c) The result of the proposed text detection method

Fig. 1. Illustrating the challenges of arbitrarily-oriented text detection in low light images through enhancement.

Fig. 1(c) shows the text detection results of our proposed method, which can be seen to work well for arbitrarily-oriented text in low light images.

II. Related Work

The scope of the method we describe encompasses image enhancement and text detection for low light images. Here we review existing approaches that attempt to address related problems.

A. Image Enhancement

Dong *et al.* [15] proposed a fast, efficient algorithm for the enhancement of low light videos. The method explores the idea of de-hazing for enhancing details. This method works well for images that contain sky regions or open environments but not for images that contain indoor scenes as background. Jiang *et al.* [1] proposed night video enhancement using an improved dark channel prior. The method uses the improved dark channel prior model, and integrates it with local smoothing and image Gaussian pyramid operators. However, this method is limited to images of outdoor environments. Sharma *et al.* [16] proposed contrast enhancement using pixel-based image fusion in the wavelet domain. The method proposes wavelet decomposition to obtain high frequency sub-bands. The method is good for images that contain the uniform contrast but not images affected by night light illumination. Rui and Guoyu [17] proposed a medical X-ray image enhancement method based on TV-Homomorphic filter. However, the method is limited to specific X-ray images enhancement. Ravishankar *et al.* [18] proposed acoustic image enhancement using Gaussian and Laplacian pyramid. However, the scope of the method is aimed at noisy images and not low light images.

From the above discussion, it is noted that most past methods are proposed for enhancing low contrast and low visibility of general images, but not images containing text. In addition, sometimes the methods target specific applications for enhancement. Hence, they may not perform well for low light text images. With the same objective, Raghunandan *et al.* [19] proposed a Riesz fractional based model for enhancing license plate detection and recognition. However, the scope of the method is confined to low contrast license plate images. Zhang *et al.* [20] proposed a new fusion-based enhancement for text detection in night video footage. However, the method does not work well for complex images although it is proposed for low light text images due to the condition used in fusion operation. It is noted from the above enhancement methods that though these methods are developed for enhancing fine details and low contrast texts in images, they do not give satisfactory results especially for low light arbitrarily-oriented text images.

B. Text Detection in Natural Scene Images

Gao *et al.* [21] proposed reading scene text with fully convolutional sequence modeling. Their method explores a stacked convolutional neural network to overcome the problems of traditional bi-directional long short-term memory. Deng *et al.* [22] proposed detecting text of different orientations with corner-based region proposals. In the first stage, the method finds possible locations for text instances based on linking corners, and in the second stage, the method explores pooling layers to fix bounding boxes for text lines of any orientation. Xue *et al.* [23] proposed multi-scale shape regression for scene text detection. The method finds dense text boundary points to overcome the weakness of conventional networks especially in handling curved text lines. Ma *et al.* [24] proposed arbitrarily-oriented scene text detection via rotation proposals. The method explores a network for extracting direction of text, and the rotation region of interest pooling is used to fix bounding box of each text line of any orientation. Bazazaian *et al.* [25] proposed facilitated and accurate scene text proposals through FCN guided pruning. The method aims at combining text proposals and fully convolutional neural networks to reduce the number of proposals that do not contribute for text detection, which results in robust results. Liu *et al.* [26] proposed curved scene text detection via transverse and longitudinal sequence connection. The method attempts to introduce context information by integrating recurrent transverse and longitudinal offset connections. Tian *et al.* [27] proposed text flow, which includes a unified text detection system for natural scene images. The main objective of the work is to combine several stages of text detection as one step to minimize errors such that text detection performance improves. Xue *et al.* [28] introduced border semantic awareness and bootstrapping for text detection in naturals scene images. The approach finds relationship between word or text lines using borders of texts. Van *et al.* [29] proposed a pooling based method for text detection in natural scene images. Their score function is designed based on histogram oriented features, which help the method to determine ranking of proposals. However, the methods may not

perform well for the low and limited light images because this constraint has not been considered in these works.

Liao *et al.* [30] proposed a method for text detection in natural scene images based on rotation-sensitive regression (RRD). The approach proposes two kinds of network, one for extracting rotation-sensitive and another for rotation-insensitive features. However, the method does not work well for images having dense text lines. Liao *et al.* [31] proposed another method based on single shot-oriented scene text detector for natural scene images (TextBoxes++). The main target of this work is to develop a single pass network for detecting text in images without using any post-processing steps to improve performance. However, the method is limited to text lines with the uniform spacing. Lyu *et al.* [32] proposed an approach for text detection in natural scene images based on corner localization and region segmentation (CLR). Corners are used to localize text boundaries and relative positions are used for text region segmentation.

The method also suffers from the same problem of non-uniform spacing of text lines. In another work, Lyu *et al.* [33] used an end-to-end trainable network for text spotting in natural scene images. The approach extracts semantic information of text to segment text regions using a deep learning model (TextSpotter). However, the method may not work well for images with dense text lines. Long *et al.* [34] proposed a method for arbitrary shaped text detection in natural scene images using flexible representation (TextSnake). The approach finds the relationship between text instances, and estimates symmetry axis by proposing a deep learning model. However, the method is considered as computationally expensive.

It is also found from the above review that the main objective of existing methods is to achieve better results for natural scene images. In addition, the primary goal of these methods is to overcome the weakness of deep learning models such that text detection performance can be improved. However, sometimes, this idea may not give consistent results for text in different type images. For example, to addresses challenges of text detection in multi-type images, such as video, natural scene and license plate images, Shivakumara *et al.* [35] proposed a fractional means-based method for multi-oriented keyword spotting in different type images. The context feature is used for text detection in multi-type images. However, the scope of the method is still limited to daytime images and not low light images.

Recently, Xue *et al.* [36] proposed curved text detection in blurred and non-blurred natural scene images. The method explores gradient information in a different way for addressing the challenges of blurred and non-blurred natural scene images. However, the method focuses on images affected by only blur information but not those affected by low light and night effects as the proposed work. Therefore, based on the above review, one can conclude that the task of detecting arbitrarily-oriented text in low light images is challenging and still an open problem.

C. Text Detection in Low/Limited Light Images

Huang *et al.* [37] proposed end-to-end vessel plate number detection and recognition using deep convolutional neural networks and LSTMs. The deep learning model has been tuned for achieving results irrespective of light conditions. As a result, the method is good for number detection in ship vessels. Panahi *et al.* [38] proposed accurate detection and recognition of dirty vehicle plate numbers for high-speed applications. The method uses adaptive thresholds for segmenting regions of interest as license plate numbers, and then uses conventional classifiers for detecting license plates in images. However, the method focuses on license plate detection, but not more general kinds of text in natural scene images. Wahyono and Jo [39] proposed LED dot matrix text recognition in natural scenes. The method uses Canny edge operator for obtaining edge components. This method works well for specific datasets and applications. Shemarry *et al.* [40] proposed ensemble of AdaBoost cascades of 3L-LBPs (Local Binary Patten) classifiers for license plates detection with low quality images. The method explores LBP features for detecting license plates, which include low light images. The success of the method depends on its preprocessing steps.

Lin *et al.* [41] proposed an efficient license plate recognition system using convolution neural networks. The method detects vehicle first, and then retrieves license plate numbers of vehicles. However, the method is not robust because it requires car shapes. Mohanty *et al.* [42] proposed an efficient system for hazy scene text detection using deep CNN and patch NMS. In this work, the method considers scenes affected by haze, smoke, and smog as poor-quality images. The performance of the method depends on the classification of hazy scene images. In this past work, methods have been developed to deal with low or limited light images, but they are generally tuned to work for a specific category of input. Furthermore, the license plate methods work well for text in a horizontal direction, but not arbitrarily-oriented text. Therefore, one can conclude that none of the methods focuses on the challenges of low light images.

Here we present new work designed for text detection in low light or limited light natural scene images. The proposed method comprises image enhancement and text detection. Inspired by the method [20] where the combination of spatial domain and frequency domain-based features are used for enhancing text information in night images, in this work we explore the same combination in a novel way, including a new fusion operation for low light image enhancement. The main basis for proposing the combination is that for smooth regions, intensity values are lower than non-smooth regions where there are edges. Similarly, the homomorphic filter in frequency domain reduces low frequency and increases high frequency information. Thus it can reduce illumination change and sharpen edges or details. Since input low light images are affected by both poor quality and loss of visibility, we believe that the above combination can cope with the challenges. Additionally, motivated by the method [43] where Cloud of Line Distribution (COLD) features are proposed to cope with variations of handwriting, we propose COLD features for handling causes of low light images in this work. Similarly, it is noted from the Related Work Section that deep learning models have been successfully applied to other complex computer vision problems and hence we explore the deep learning model for text detection in this work by feeding COLD feature to convolutional neural networks.

We summarize the contributions of the proposed method as follows. Although there has been some past work on image enhancement and text detection in poor quality images, the methods either focus on spatial domain-based features or frequency domain-based features. None of them proposes a combination of the two kinds of features for enhancing fine details in low light natural scene text images. In the same way, COLD has been used for handwriting analysis in document images, but not for text detection in low light images. Instead of using the features extracted from input images directly, the proposed approach combines handcrafted features with deep ones in a new way to improve arbitrarily-oriented text detection performance in low light images.

III. Proposed Method

Since the proposed method aims at text detection in low light or limited light natural scene images with arbitrarily-oriented texts, for each input image our method first explores image enhancement for improving image quality. This step makes edges visible, and ensures that edges preserve structures of characters. It is true that high intensity values represent edges in spatial domain and high frequency coefficients represent edges in frequency domain. In order to take the advantages of these two observations, we explore the combination of spatial and frequency-based features through fusion for enhancing fine details in images. For each enhanced image, we propose to use the Maximally Stable Extremal Regions (MSER) step for detecting text candidates. This is because MSER exploits the fact that text pixels in the same image share almost the uniform values, and it is not sensitive to arbitrary orientations. Sometimes, the step of image enhancement may also enhance objects in background and the same detected by MSER as text candidates. This motivated us to further propose the combination of COLD features (cloud of line distribution) and deep convolutional neural networks.

The reason to propose COLD is that it helps to characterize stroke direction and behavior in handwriting for writer identification. The same stroke characteristics can be used for separating objects in the background from text. Thus, we employ COLD for detecting false text candidates. In the same way, deep convolutional neural networks exhibits a strong discriminative power, thus we use them to strengthen the COLD features for false text candidates. This step outputs text components. In order to fix bounding boxes for arbitrarily-oriented text lines, the proposed method performs grouping operation based on the direction and proximity between characters. This results in text detection for low light images of any orientation. The whole block diagram can be seen in Fig. 2.

For the images captured in low light, it is difficult to read the content (for example, see the sample low light images shown in Fig. 3(a)). However, when we look at pixel values, there is a difference between edge and background pixels. This observation motivates us to propose an enhancement method in this work. To justify this observation, we employ clustering that chooses the maximum and minimum values from input image. Then it classifies the values which are close to the maximum as a Max cluster and otherwise as Min cluster. It is expected that the Max cluster contains high values that represent edge information. The pixels in the Max cluster are displayed as white pixels for the input as shown in Fig. 3(b), where it is noted that text information is visible compared to the input image.

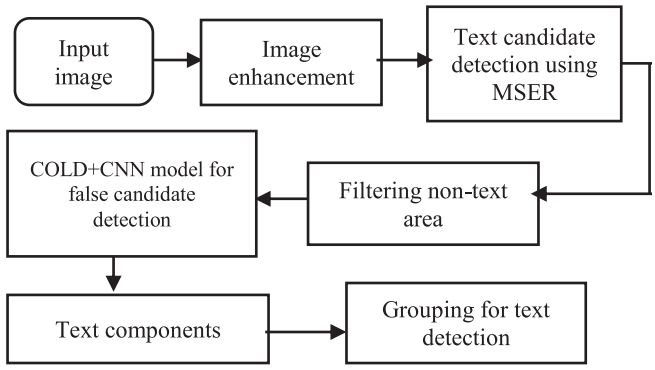


Fig. 2. The architecture of the proposed method.

(a) Input images (b) Max cluster results

Fig. 3. Indicating there is a difference between edge and background pixels in the low light mages.

A. Fusion Method for Image Enhancement

Inspired by the method [15] where atmospheric light and intensity values for dehazing are used, we explore the concept in a different way for enhancing fine details of texts in low light images. The proposed method performs inverse operation to obtain an invert image for the input gray image. The method estimates percent of light emitted from objects or scene in the image based on based on the distance between pixels and camera. Values of percent are used to study smooth and rough regions in the image. The proposed method also uses adaptive thresholds for determining pixels that represent smooth and heterogeneous regions. In general, we believe that smooth and heterogeneous regions indicate dark color that does not contain text information and edges that contain text information, respectively. The effect of the above process is illustrated in Fig. 4, where (a) denotes input low light images, and (b) gives the results of the above process, which is called the spatial domain-based enhancement method. It is noted from the results shown in Fig. 4(b) that one can read the text.

Similarly, inspired by the method [17] where homomorphic filter is proposed for enhancing poor quality of x-ray images, we explore the concept in a different way for enhancing low light images in this work. The proposed method considers image enhancement as a restoration problem as defined in Equation (1), which is called the total variation model.

$$u\ (x,y) = u_0\ (x,y) + n\,(x,y) \tag{1}$$

(a) Input low light images

(b) The result of spatial domain based enhancement method

(c) The result of frequency domain based enhancement method

(d) The result of the proposed fusion method

Fig. 4. The effect of the proposed fusion method for enhancing low light images.

where $u(x,y)$denotes an image with noise,$u_0(x,y)$ denotes an image without noise, and $n(x,y)$ represents noises. The Equation (1) shows the total variation model is a nonlinear filter, which can be used for enhancing $u_0(x,y)$by suppressing $n(x,y)$. Therefore, for an image $n(x,y)$, Equation (1) can be modified for $n(x,y)$ as defined in Equation (2).

$$\mathrm{TV}\ (u) = \int_{\Omega} |\nabla_u|\, d\Omega \tag{2}$$

where $|\nabla_u| = \sqrt{u_x^2 + u_y^2}$, and Ω represents the boundary of the image $u(x,y)$. Noise reduction of an image is converted to the process of minimizing the total variation, and the problem of minimizing the constraint of the total variation is defined in Equation (3)

$$\min_u \int_{\Omega} |\nabla_u| + \frac{1}{2} u_0 - u_2^2 \tag{3}$$

Equation (3) is solved using the Euler-Lagrange formula as defined in Equation (4), where the gradient descent step is used. This results in a denoised image as shown in Fig. 3(c), where it can be noticed that the details are enhanced compared to the input ones in Fig. 3(a).

$$\lambda\,(u_0 - u) + div\ (\nabla_u / |\nabla_u|) = 0 \tag{4}$$

It is also observed from Fig. 3(b) and Fig. 3(c) that the spatial domain based and frequency domain-based enhancement steps magnify the details in low light images. Since the nature of low light images is unpredictable, an individual enhancement step may not produce consistent results for all situations. To alleviate this problem, we further propose a fusion operation to integrate the results of the spatial and frequency domains based steps. For fusing, the proposed method calculates the mean and standard deviation for each image as defined in Equation (4) and Equation (6).

$$m^j\ (x,y) = \frac{1}{n*n} \sum_{x\,=\,0}^{n-1} \sum_{y\,=\,0}^{n-1} f^j\,(x,y) \tag{5}$$

$$s^j\ (x,y) = \sqrt{\frac{1}{n*n} \sum\nolimits_{x=0}^{n-1} \sum\nolimits_{y=0}^{n-1} \left[f^j\,(x,y) - m^j\,(x,y)\right]^2} \tag{6}$$

where$j$ denotes image index and j$\epsilon\{1,2\}$, $m^j(x,y)$ denotes the mean value, $s^j(x,y)$ denotes the standard deviation, $f^j(x,y)$ indicates the pixel value of the image with x row y column, and n is the size of the window. For *n*, we determine it empirically as 7. The proposed method derives weights using the mean and standard deviation as defined in Equation (7) for the results of the spatial and frequency domains based steps.

$$t^k\ (x,y) = \frac{m^k\,(x,y) * s^k\,(x,y)}{\sum_{l=1}^{p} m^k\,(x,y) * s^k\,(x,y)} \tag{7}$$

where $p\ =\ 7, k\epsilon\{1,2\}$. denotes the label of the image To tune weights for better enhancement, the proposed method redefines weights as defined in Equation (8)-Equation (10).

$$c\ (x,y) = \sqrt{\sum\nolimits_{l=1}^{n} \left\{t^l\,(x,y)\right\}^2} \tag{8}$$

$$w^1\ (x,y) = \frac{1}{\sqrt{1 + \exp\left(-c\,(x,y)\right)}}\, w \tag{9}$$

$$w^2\ (x,y) = 1 - w^1\,(x,y) \tag{10}$$

where in Equation (8), $n = 2$, $w^1(x,y)$ denotes pixel weight of the enhanced image given by the spatial domain based method $f^1(x,y)$, and $w^2(x,y)$ is the weight of the enhanced image given by the frequency domain based method $f^2(x,y)$. The fusion operation using spatial and frequency domains based enhanced images is defined as Equation (11).

$$F\ (x,y) = \sum_{q=1}^{n} w^q\ (x,y) \cdot f^q\,(x,y) \tag{11}$$

The effect of the fusion can be seen in Fig. 4(d), where one can see texts are visible clearly compared to the input low light images in Fig. 4(a). This is the advantage of the proposed fusion method.

B. Arbitrarily-Oriented Text Detection in Low Light Natural Scene Images

Use of MSER for text detection in natural scene images is studied widely [44], [45]. This is the motivation for us to explore MSER for text candidate detection in the proposed work. It is observed from the existing methods that most of the methods

Fig. 5. MSER for text candidate text detection using different color spaces.

Fig. 6. Text candidate detection using MSER for the enhanced images.

ignore color information for text detection. In our case, we employ MSER on R, G and B color spaces of the enhanced images as shown in Fig. 5, where one can see MSER detects almost all the text components in color spaces along with background information, which are called text candidates. However, the fusion operation and connected component analysis remove non-text candidates as shown in Fusion in Fig. 5, where it is noted that almost all the non-text candidates are removed.

This is valid because the enhancement step restores character structures in text lines and widens the gap between background and foreground (text). The effect of MSER with fusion on the input images shown in Fig. 4(d) can be seen in Fig. 6(a), where we can see bounding boxes for the text candidates given by MSER. The text candidates are displayed in a binary form as shown in Fig. 6(b). Note that from Fig. 6(b), it is also found that MSER detects non-texts as text candidates due to background objects which share the properties of characters.

It is true that eliminating false text candidates is hard due to common features shared by text and background objects. Motivated by the method [43] where COLD is used for writer identification irrespective of handwriting variation, we explore the concept in a different way to cope with the challenges of text and non-text separation in this work. For each text candidate given by MSER, the COLD finds contour points with the help of polygonal approximation. For each contour point, the COLD estimates the distance between the nearest contour points in polar domain, which results in distribution of points as shown in Fig. 7(a) and Fig. 7(b) for text and non-text candidates, respectively. Fig. 7 shows the COLD distribution for distance k $=$ 3. Different values of k are considered for feature extraction. It is observed from Fig. 7 that for texts, the COLD distribution appears dense, while for non-texts, the distribution appears scattered. As the number of straight segments increases, the density

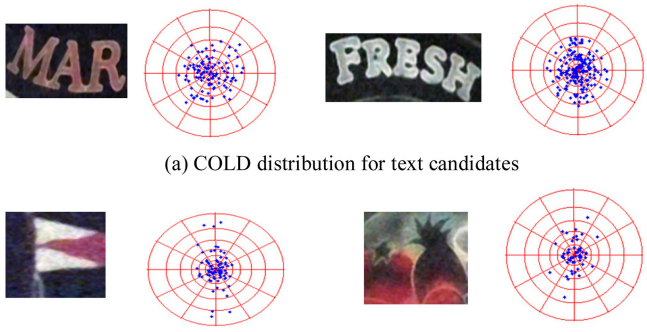


Fig. 7. COLD distribution for text and non-text candidates.

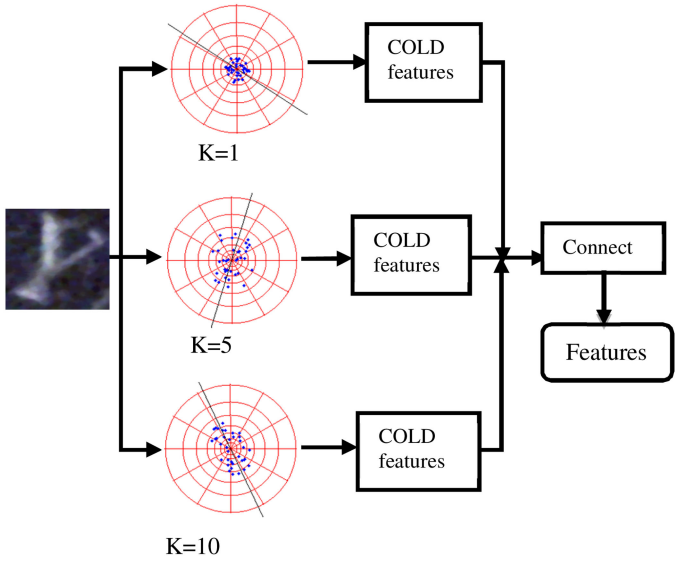


Fig. 8. COLD features extraction for text candidates.

of points increases, otherwise the density decreases. In general, texts involve more straight segments compared non-texts in almost every case. Therefore, COLD gives dense distributions for text candidates and scattered distributions for non-text candidates.

To extract such properties of COLD, the proposed method draws rings over the distribution with fixed radius as shown in Fig. 8. After drawing rings, the proposed method divides the whole distribution into segments, which can be labeled as sectors including 7 rings with 12 angles. This results in 84 segments like boxes as shown in Fig. 8. The proposed method counts the number of points in each segment, which are considered as density features. For the values of different k ($k = 1, k = 5$, and $k = 10$), the proposed method extracts density features, which results in a 252-dimensional vector as shown in Fig. 8. These features are fed to the deep learning model for eliminating false text candidates, which result in text detection.

As mentioned in the Proposed Methodology Section, deep learning models haven been successful for solving complex problems, we thus propose to integrate COLD features with deep features together for eliminating false text candidates. The architecture of integrating COLD and deep features can be seen in

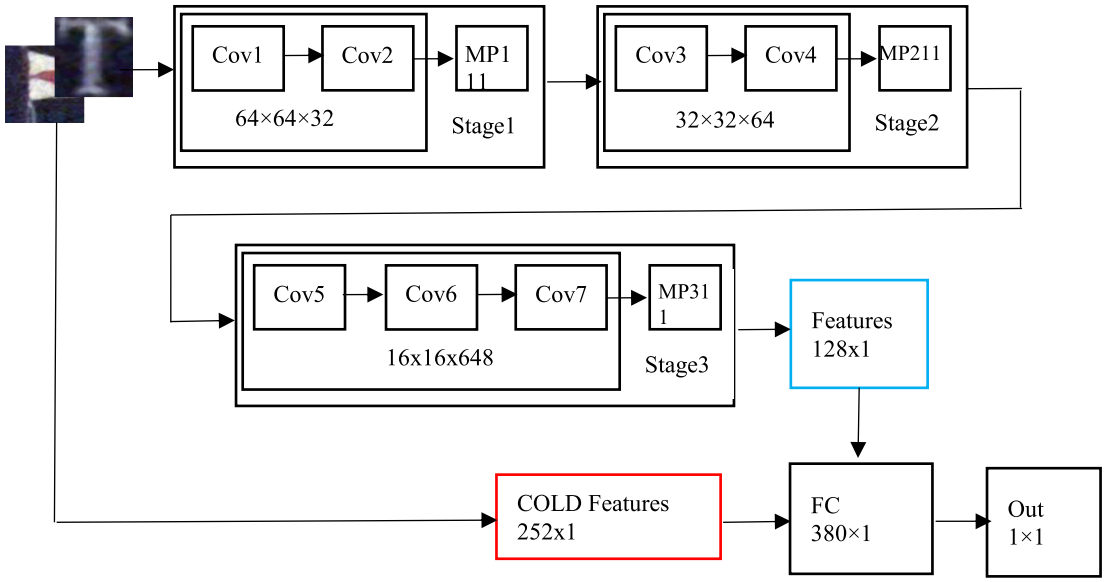
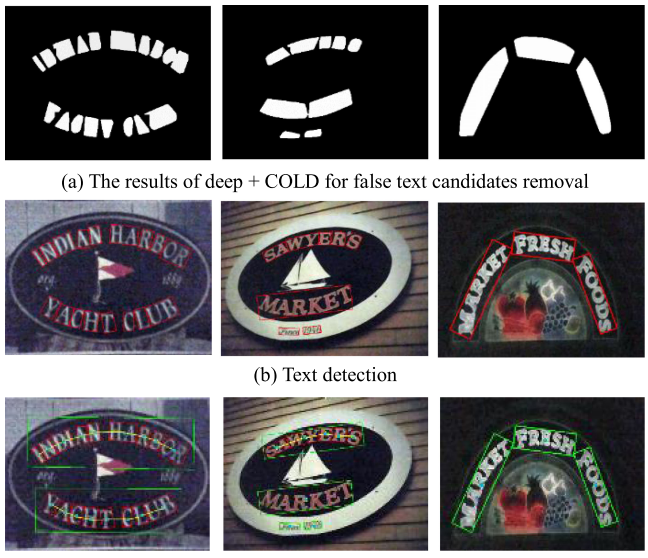


Fig. 9. The architecture for combining the deep and COLD features. The convolutional, max-pooling and full-connect layers are denoted by COV, MP, and FC respectively. The blue color indicates the deep features and the red color indicates the COLD features.

Fig. 9, where we add COLD as well as deep features. For training and learning, we use text candidates detected by MSER step from ICDAR 2013, ICDAR 2015 and our data. Based on the results, we label text and non-text candidates. Further, to improve the robustness of the classification of text and non-text candidates, we rotate text and non-text candidates randomly, which results in 50000 non-text candidates and 50000 text candidates, which are considered as positive samples. In this way, the proposed deep learning model extracts a 128-dimensional feature vector.

Fig. 9 shows that the architecture requires a 3-channel candidate of size 64×64, namely, the first and second convolution layers (COV1, COV2) consider 32 kernels of size $3 \times 3 \times 1$ with a padding of 1 pixel. After convolving with multiple filter masks, the proposed architecture deploys Rectified Linear units (ReLU) as the non-linear activation function, which results in a $64 \times 64 \times 32$ feature dimension matrix. This feature matrix is then fed to Max-Pooling Layer (MP1), which fuses 2×2 spatial neighborhoods with a stride of 2 pixels. In fact, the convolutional layers of the proposed architecture are the core which provides various and hierarchy feature maps of appearance, while the max-pooling layers offer activation features with the ability of robustness to slight shifts in appearance. The output of the third max-pooling layer is 128 dimensions deep convolutional feature.

Next, we integrate the COLD features with the deep features, which are then feed to the fully-connected layer (FC) to assign the binary label, namely, text or non-text for input text candidates. This outputs text detection results by removing false text candidates as shown in Fig. 10(a). Fig. 10(b) shows text detection and further text components are grouped at the component level based on both directions of text lines and proximity between characters as shown in Fig. 10(c), which are correct results of text detection of arbitrarily-oriented texts in low light images.



(a) The results of deep + COLD for false text candidates removal

(b) Text detection

(c) The bounding boxes are fixed for arbitrarily-oriented text lines

Fig. 10. The proposed Deep + CNN for text detection in the low light images.

Since the proposed method obtains character components with the help of the combination of MSER, deep learning models and COLD, we use boundary points of components to fix bounding boxes for arbitrarily-oriented text lines [27]– [29] as illustrated in Fig. 10(a)–Fig. 10(c). For all of our evaluations, we use 75% of the dataset for training and 25% for testing. In the case of the benchmark datasets, we follow the standard evaluation schemes employed by previous researchers.

IV. Experimental Results

Since there are no existing datasets for text detection in low-light images, we create our own for experimentation. For

Fig. 11. Sample images for each dataset considered for experimentation.

TABLE I
DETAILS OF THE DATASETS CONSIDERED FOR EXPERIMENTATION (LR: LOW RESOLUTION, OR: ORIENTATION, ML: MULTI-LINGUAL, AOC: ARBITRARY ORIENTED/CURVED, LV: LOSS OF VISIBILITY AND LC: LOW CONTRAST)

Datasets	Train	Test	LR	OR	ML	AOC	LV	LC
Our data	300	200	√	√	√	√	√	√
ICDAR2013	258	251	-	-	-	-	-	-
ICDAR2015	1000	500	√	√	-	-	-	-
SVT	100	250	-	√	-	-	-	-
Total-Text	1225	300	-	√	-	-	-	-
ICDAR-MLT2017	7200	900	-	√	√	√	-	-
CTW1500	1000	500	-	√	√	√	-	-

TABLE II
ANALYZING THE QUALITY OF THE TEXT IN OUR AND STANDARD DATASETS

Datasets	BRISQUE	NIQE	PIQE
Our data	53.12	9.79	72.65
ICDAR2013	41.21	16.29	56.22
ICDAR2015	43.18	18.3	67.21
SVT	47.09	40.6	79.38
Total-Text	41.74	11.32	48.97
ICDAR2017	40.33	15.85	60.91
CTW1500	38.61	15.7	49.94

evaluating the proposed enhancement step, we employ standard quality measures. To test the robustness of the proposed method beyond low-light images, we use benchmark natural scene text datasets for experimentation. Furthermore, discussions on effectiveness and usefulness of the proposed method comparing with the existing methods based on experimental results are analysis and presented.

A. Dataset Creation and Evaluation

Our dataset includes images taken in low light, moonlight, and street lamp light, which introduce low contrast and resolution and the images with complex background, arbitrarily-oriented, and multi-lingual texts. In total, the dataset contains 500 images for experimentation. Out of 500 images, as per the standard norms for choosing samples for training and testing, we use 75% of the images for training and 25% for testing. If we count the number of text instances in images, it is somewhat more than 1200 text instances. The main challenge of this dataset is the loss of visibility or contrast as the shown sample images in Fig. 11. For collecting the dataset, we used several mobile cameras of different capacities by varying distances from camera and scenes to include more variations in the dataset. The dataset can be found by [46].

In order to test the robustness of the proposed method, we also consider the standard datasets of natural scene images, namely, ICDAR 2013, ICDAR 2015, SVT, Total-Text, ICDAR2017-MLT and CTW1500 datasets, for experimentation. Sample images for respective of datasets are shown in Fig. 9, where we can see the complexity of text detection varies from one dataset to another. The ICDAR 2013 and ICDAR 2015 datasets are simple compared the other datasets because these two datasets do not contain many curved text images and other variations. The SVT dataset consists of street view images, and hence these images have complex background. The Total-Text and CTW1500 datasets are more complex compared the other datasets because of curved texts with complex background. However, the ICDAR2017-MLT dataset provides images of multi-lingual text with orientations and complex background. In summary, the natures of respective datasets are listed in Table I.

For measuring the performance of the proposed method that involves the enhancement and text detection steps, we consider the standard quality measures, namely, BRISQUE, NIQE and PIQE, for the enhancement step, and Recall (R), Precision (P) and F-measure (F) for the text detection step. Here BRISQUE measures spatial quality [47], NIQE measures naturalness of images [48], while PIQE measures image quality [49]. Low values of all the three measures indicate better perceptual quality. We prefer to choose these measures for evaluating the enhancement step because these measures do not require ground truth for calculating values. However, PSNR, MSE and SSIM require ground truth for estimating values. In addition, the measures are available in the form of in-built functions. For calculating these measures, we follow the instructions mentioned in [8] about the evaluation scheme for all the experiments. In case of our data, since there is no ground truth, we manually count the measures, while for the other standard datasets, ground truth helps us to calculate measures, automatically.

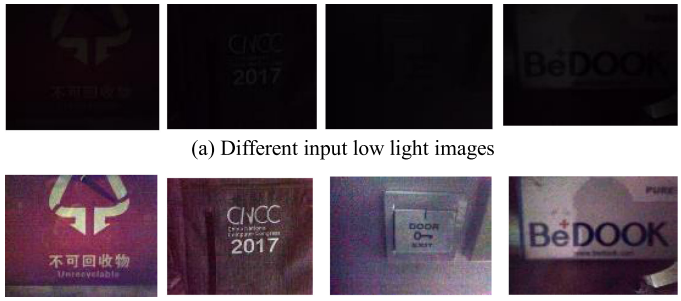
To analyze the quality of the text in our and the standard datasets, we estimate quality measures as reported in Table II, where BRISQUE scores a high value for our dataset compared to all the other datasets. This indicates that our dataset has more poor quality images due to low light and limited light. At the same time, NIQE and PIQE measures give the highest values for the SVT dataset compared to all the other datasets including

ours, which shows SVT dataset has more poor-quality images caused by distortion. However, when we look at the scores of the measures of ICDAR 2013, ICDAR 2015, Total-Text, ICDAR 2017 MLT and CTW1500 datasets, the measures do not have a significant difference. Therefore, we can infer that respective datasets include poor quality images either caused by distortion, low resolution or low contrast. However, if we analyze it more deeply, it can also be seen from Table II that for ICDAR2015 dataset, BRISQUE, NIQE and PIQE values are higher compared to those of ICDAR2013, Total-Text, ICDAR 2017 MLT and CTW1500. This indicates that ICDAR2015 dataset has more-lower quality images than ICDAR2013, Total-Text, ICDAR2017 MLT and CTW1500.

To show the effect of the enhancement step and the proposed text detection step, we implement two enhancement methods, namely, Dong *et al.*'s method [15] which proposed a fast efficient algorithm for the enhancement of low light videos, and Rui and Guoyu's method [17] which proposed a medical x-ray image enhancement method based on TV-homomorphic filter. Since there is no specific method for enhancing low contrast texts in low light images, we prefer to use the above-two mentioned methods for comparative studies. The reason is that the method in [15] focuses on low light videos and the method in [17] focuses on poor quality images as the proposed work. Similarly, for text detection, we use the code of reproducibility given by the methods consisting of Zhou *et al.*'s [5] which proposed an efficient and accurate scene text detector (EAST), Shi *et al.*'s [6] which proposed detecting oriented texts in natural scene images by linking segments (SegLink), and Tian *et al.*'s [50] which proposed detecting texts in natural scene images with connectionist text proposal networks (CTPN). The reason to choose the above methods is that they addressed challenges of multi-orientation, low contrast, low resolution, complex background, etc. In addition, their codes are available publicly for experimentation. The method in [36], which is developed for curved text detection in blurred and non-blurred natural scene images, is also considered for comparative study with the proposed method. This method attempts to address the issue of blur in images, which is also one of the challenges we consider in our work. The methods are run on all the datasets for experimentation in this work. Apart from that, for the standard datasets, we report the results as mentioned in the papers for comparative studies.

Note that all the datasets include images captured in daylight while our dataset include images captured in low light conditions. Therefore, we conduct experiments before enhancement and after enhancement to test the effectiveness of the proposed enhancement step. In other words, we run different text detection methods on images of our dataset without enhancement to calculate the measures. In the same way, we run different text detection methods on enhanced images to calculate the measures. It is expected that the text detection performance will be lower before enhancement compared to after enhancement.

Since our objective is to show the enhancement is effective for improving the performances of text detection methods on low light images, the proposed method does not use enhanced images for learning to calculate measures. For learning, we use pre-defined samples of input images according to the datasets.



(a) Different input low light images

(b) The results of the proposed enhancement

Fig. 12. Examples of enhancement of the proposed steps for different low light input images.

TABLE III
QUALITY MEASURES FOR THE INPUT AND ENHANCED IMAGES OF THE PROPOSED AND EXISTING METHODS

Methods	BRISQUE	NIQE	PIQE
Input Low Light Images	49.23	4.38	32.49
Dong et al. [15]	43.52	4.12	26.53
Rui et al. [17]	41.23	3.94	23.27
Proposed Enhancement	38.51	3.76	20.86

The same set up is used for all the experiments in this work. The reason is that if we use enhanced images for learning parameters to calculate performance measures, one cannot see the effectiveness of the enhancement comparing to input images.

B. Evaluating Enhancement

Qualitative results of our method for different input low light images are shown in Fig. 12, where it is noted that the proposed approach enhances the details such that we can read the content in images. Note that one cannot read the content in the original input images directly. Fig. 12 shows that our method enhances fine details in the images irrespective of text orientation, contrast, resolution and background. The main reason to achieve good results is that the methods fuse merits of spatial and frequency domains. Quantitative results of the proposed and existing methods for the enhancement are reported in Table III, where it is noted that all the three quality measures score low values for the proposed approach compared to the existing methods. As mentioned in the previous section, low values of the measures indicate better perceptual quality. Therefore, one can assert that our method preserves better spatial, naturalness and image quality compared to the existing methods. The reason for the poor results by the existing methods is that these methods are developed for particular cases but not for images affected by both poor quality and low contrast. On the other hand, the proposed approach is better than the existing methods because of the new fusion to integrate the strengths of spatial and frequency features.

In the proposed work, feature extraction using COLD is a key step to achieve better results for text detection in low light images. To analyze the effectiveness of COLD, we calculate measures for input and enhanced images of all the datasets including our dataset before and after enhancement as reported

TABLE IV
COMPARISONS OF THE PROPOSED METHOD WITHOUT COLD AND WITH COLD (RECALL-R, PRECISION-P, AND F-MEASURE-F).

Datasets	Proposed (without COLD)						Proposed (with COLD)					
	Before Enhancement			After Enhancement			Before Enhancement			After Enhancement		
	P	R	F	P	R	F	P	R	F	P	R	F
Our Dataset	0.48	0.27	0.35	0.53	0.47	0.49	0.60	0.31	0.40	0.65	0.56	0.60
ICDAR 1013	0.79	0.61	0.69	0.79	0.61	0.69	0.84	0.78	0.81	0.84	0.78	0.81
ICDAR 2015	0.67	0.43	0.52	0.67	0.43	0.52	0.70	0.48	0.57	0.70	0.48	0.57
SVT	0.73	0.67	0.70	0.72	0.66	0.69	0.77	0.69	0.73	0.76	0.68	0.72
Total-Text	0.63	0.42	0.50	0.62	0.42	0.50	0.67	0.43	0.52	0.66	0.43	0.52
ICDAR 2017 MLT	0.59	0.46	0.52	0.59	0.46	0.52	0.63	0.49	0.55	0.63	0.48	0.54
CTW1500	0.52	0.49	0.50	0.52	0.49	0.50	0.56	0.51	0.53	0.56	0.51	0.53

(a) Text detection of the proposed method for the input low light images

(b) Text detection of the proposed method for the enhanced images.

Fig. 13. Examples of text detection of the proposed method on low light images.

in Table IV. For this experiment, the proposed method extracts features from text candidates given by MSER step and input images, and then fed them to a deep learning network for text detection. The results of the proposed method with COLD and without COLD are reported in Table IV, where it is noted that the proposed method without COLD achieves poor results for all the datasets compared to the proposed method with COLD. Therefore, one can argue that features extracted from COLD distribution has good contribution for achieving better results.

C. Evaluating Text Detection on Our Low Light Image Dataset

Qualitative results of the proposed text detection on input images without enhancement and the output of the proposed enhancement step are shown in Fig. 13(a) and Fig. 13(b), where it can be seen that the proposed text detection works well for enhanced images but misses texts for the original input images. It is expected because of loss of visibility and contrast in case of low-light input images. This shows that for the input images shown in Fig. 12(a), enhancement is essential to improve visibility and quality of images. Quantitative results of the proposed and existing methods for the input images and enhanced images are reported in Table V, where we can find that all the methods including the proposed method scores better results for enhanced images (after enhancement) compared to input ones (before enhancement). There is a significant improvement in

TABLE V
TEXT DETECTION PERFORMANCE OF THE PROPOSED AND EXISTING METHODS BEFORE ENHANCEMENT AND AFTER ENHANCEMENT ON OUR DATASET

Methods	Low-light images (Input) (before enhancement)			Enhanced images (output) (After enhancement)		
	Precision	Recall	F-measure	Precision	Recall	F-measure
EAST [5]	**0.73**	0.18	0.30	0.70	0.51	0.59
SegLink [6]	0.43	0.16	0.23	0.47	0.37	0.42
CTPN [50]	0.67	0.20	0.31	0.69	0.43	0.53
TextSnake[9]	0.63	0.26	0.37	0.74	0.53	0.62
TextField[31]	0.66	0.27	0.38	**0.78**	0.55	**0.63**
Xue et al.[36]	0.57	0.17	0.26	0.63	0.52	0.57
Proposed	0.60	**0.31**	**0.40**	0.65	**0.56**	0.60

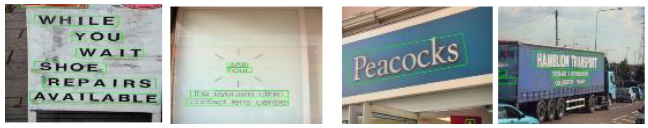
the performances of the proposed and existing methods for enhanced images compared to before enhancement especially in terms of F-measure. This indicates that the proposed enhancement plays a vital role in enhancing the performances of text detection methods for low-light images.

Table V shows that the proposed method achieves the best F-measure for low light images (before enhancement) and Recall for enhanced images (after enhancement) compared to the existing methods. The EAST method scores the higher Precision for low light images, but reports the lower Recall compared to the proposed method. Similarly, the TextField method is the best at Precision and F-measure for enhanced images compared to the other methods including the proposed method. This infers the deep learning state-of-the-art methods detect texts well for enhanced images. It is also observed from Table V that all the methods including the proposed method score low recall compared to precision for the input images. The methods miss some text information due to the loss of contrast, while for enhanced images the Recall improves significantly for all the methods. The existing methods report poor recall for enhanced images compared with the proposed method. This is due to inherent limitations of the methods to handle the causes of low light images. On the other hand, since the objective of the proposed method is to detect arbitrary texts in low light images, the proposed method achieves the best F-measure for input images and Recall for enhanced images.

D. Evaluating Text Detection on Benchmark Natural Scene Datasets

As noted from the previous section, the proposed method has the ability to handle both low light images and good images. To validate the performance of the proposed method, we conduct experiments on the benchmark datasets, namely, ICDAR 2013, ICDAR 2015, SVT, Total-Text, ICDAR 2017-MLT and CTW1500, where there are no low light images. In other words, it is expected that the proposed method should report consistent results for input images (before enhancement) and enhanced images (after enhancement) compared to the existing methods. It is evident from the qualitative results of the proposed method (the qualitative results of ICDAR 2013, ICDAR 2015, SVT, Total-Text, ICDAR2017-MLT and CTW1500 datasets are shown respectively in Fig. 14– -Fig. 19) that the proposed method detects texts well for all the datasets.

(a) Text detection of the proposed method for the ICDAR 2013 images



(b) Text detection of the proposed method for the enhanced images.

Fig. 14. Examples of text detection of the proposed method for ICDAR 2013 natural scene dataset.

(a) Text detection of the proposed method for the ICDAR 2015 images

(b) Text detection of the proposed method for the enhanced images.

Fig. 15. Examples of text detection of the proposed method for ICDAR 2015 natural scene dataset.

(a) Text detection of the proposed method for the SVT images

(b) Text detection of the proposed method for the enhanced images.

Fig. 16. Examples of text detection of the proposed method for SVT natural scene dataset.

It is observed the results reported in Table VI -Table XI for respective ICDAR 2013, ICDAR 2015, SVT, Total-Text, ICDAR 2017-MLT and CTW1500 datasets, we can see that the results of proposed and existing methods either improve after enhancement or remain similar with the results of before enhancement. This indicates that the proposed enhancement step contributes to improve text detection performance after enhancement. This is understandable because the images of these dataset including our dataset suffer from poor quality. However, for the results of Table VI to Table XI, the proposed method reports almost the same results for input and enhanced images. But the existing methods do not score consistent results before and after enhancement for all the above-mentioned datasets. However, when we compare the results of the proposed and existing methods before and after enhancement, the existing methods score

(a) Text detection of the proposed method for the Total-Text images

(b) Text detection of the proposed method for the enhanced images.

Fig. 17. Examples of text detection of the proposed method for Total-Text natural scene dataset.

(a) Text detection of the proposed method for the ICDAR 2017-MLT images

(b) Text detection of the proposed method for the enhanced images.

Fig. 18. Examples of text detection of the proposed method for ICDAR2017-MLT natural scene dataset.

(a) Text detection of the proposed method for the CTW1500 images

(b) Text detection of the proposed method for the enhanced images.

Fig. 19. Examples of text detection of the proposed method for CTW1500 natural scene dataset.

TABLE VI
TEXT DETECTION PERFORMANCE OF THE PROPOSED AND EXISTING METHODS BEFORE ENHANCEMENT AND AFTER ENHANCEMENT ON ICDAR 2013 DATASET

Methods	Original Images			Enhancement images		
	Precision	Recall	F-measure	Precision	Recall	F-measure
EAST [5]	0.93	0.83	0.87	0.94	0.83	0.88
SegLink [6]	0.87	0.83	0.85	0.86	0.84	0.85
CTPN [50]	0.93	0.83	0.88	0.93	0.84	0.88
TextSnake[9]	0.92	0.82	0.87	0.92	0.83	0.87
TextField[34]	0.93	0.87	0.90	0.93	0.88	0.91
Xue et al.[36]	0.77	0.75	0.76	0.78	0.77	0.77
Wang et al.[51]	0.80	0.74	0.77	-	-	-
Zhu et al.[52]	0.85	0.74	0.80	-	-	-
Lu et al.[53]	0.89	0.70	0.78	-	-	-
Proposed	0.84	0.78	0.81	0.85	0.78	0.81

TABLE VII
TEXT DETECTION PERFORMANCE OF THE PROPOSED AND EXISTING METHODS BEFORE ENHANCEMENT (ON INPUT IMAGES) AND AFTER ENHANCEMENT ON ICDAR 2015 DATASET

Methods	Input Images			Enhanced Images		
	Precision	Recall	F-measure	Precision	Recall	F-measure
CTPN [50]	0.74	0.51	0.60	0.73	0.52	0.60
EAST [5]	0.78	0.83	0.80	0.76	0.84	0.80
SegLink [6]	0.73	0.77	0.75	0.72	0.75	0.74
TextSnake[9]	0.85	0.80	0.82	0.85	0.81	0.83
TextField[34]	0.84	0.84	0.84	0.85	0.85	0.85
Xue et al.[36]	0.57	0.45	0.51	0.57	0.47	0.52
NJU-Text [53]	0.70	0.36	0.47	-	-	-
AJOU [55]	0.47	0.47	0.47	-	-	-
StradVision [54]	0.53	0.46	0.50	-	-	-
SSTD[55]	0.80	0.74	0.77	-	-	-
Proposed	0.70	0.48	0.57	0.72	0.49	0.58

TABLE VIII
TEXT DETECTION PERFORMANCE OF THE PROPOSED AND EXISTING METHODS BEFORE ENHANCEMENT AND AFTER ENHANCEMENT ON SVT DATASET

Methods	Input Images			Enhanced Images		
	Precision	Recall	F-measure	Precision	Recall	F-measure
EAST [5]	0.87	0.73	0.80	0.88	0.77	0.82
SegLink [6]	0.79	0.74	0.76	0.81	0.75	0.78
CTPN [50]	0.86	0.71	0.78	0.87	0.73	0.79
Xue et al.[36]	0.71	0.58	0.64	0.74	0.60	0.67
TextSnake[9]	0.87	0.74	0.80	0.88	0.75	0.81
TextField[34]	0.85	0.79	0.82	0.86	0.79	0.83
Huang et al.[56]	0.74	0.67	0.70	-	-	-
Wu et al.[54]	0.78	0.66	0.72	-	-	-
Neumann et al.[57]	0.74	0.63	0.68	-	-	-
Proposed	0.77	0.69	0.73	0.76	0.68	0.72

TABLE IX
TEXT DETECTION PERFORMANCE OF THE PROPOSED AND EXISTING METHODS BEFORE ENHANCEMENT AND AFTER ENHANCEMENT ON TOTAL-TEXT DATASET

Methods	Original Images			Enhanced images		
	Precision	Recall	F-measure	Precision	Recall	F-measure
CRAFT [58]	0.87	0.80	0.83	-	-	-
PSENet-1s [59]	0.84	0.78	0.81	-	-	-
Textboxes [60]	0.62	0.45	0.52	-	-	-
Xue et al.[36]	0. 62	0.42	0.50	0.63	0.43	0.51
EAST [5]	0.50	0.36	0.42	0.51	0.36	0.42
Seglink [6]	0.30	0.24	0.27	0.31	0.26	0.28
CTPN [50]	0.48	0.38	0.42	0.49	0.39	0.43
TextSnake[9]	0.82	0.74	0.78	0.82	0.73	0.77
TextField[34]	0.81	0.80	0.80	0.81	0.81	0.81
Proposed	0.67	0.43	0.52	0.68	0.43	0.52

the better results compared to the proposed method. The methods TextField and TextSnake usually score the highest results for almost all the datasets before and after enhancement compared to the other existing methods and the proposed one. The reason is that these two methods are developed for addressing several challenges of natural scene text detection, such as arbitrary orientation, multi-lingual, complex background, irregular shaped text etc. But the other existing methods are inadequate to addresses the above challenges. Since the proposed method is developed for both low-light and daylight images, the proposed

TABLE X
TEXT DETECTION PERFORMANCE OF THE PROPOSED AND EXISTING METHODS BEFORE ENHANCEMENT AND AFTER ENHANCEMENT ON ICDAR 2017-MLT DATASET

Methods	Input Images			Enhanced Images		
	Precision	Recall	F-measure	Precision	Recall	F-measure
EAST [5]	0.76	0.38	0.50	0.77	0.40	0.51
SegLink [6]	0.55	0.30	0.39	0.55	0.32	0.40
CTPN [50]	0.78	0.43	0.55	0.78	0.44	0.55
TextSnake[9]	0.80	0.42	0.55	0.81	0.43	0.56
TextField[34]	0.81	0.57	0.67	0.81	0.58	0.68
Xue et al.[36]	0.63	0.17	0.27	0.65	0.19	0.29
TDN SJTU [59]	0.64	0.35	0.46	-	-	-
TH-DL [59]	0.31	0.68	0.35			
FOTS [59]	0.81	0.57	0.67	-	-	-
PSENet-1s [59]	0.77	0.68	0.75	-	-	-
Proposed	0.63	0.49	0.55	0.63	0.50	0.56

TABLE XI
TEXT DETECTION PERFORMANCE OF THE PROPOSED AND EXISTING METHODS BEFORE ENHANCEMENT AND AFTER ENHANCEMENT ON CTW1500 DATASET

Methods	Input Images			Enhanced Images		
	Precision	Recall	F-measure	Precision	Recall	F-measure
EAST [5]	0.58	0.57	0.57	0.58	0.58	0.58
Seglink [6]	0.42	0.40	0.41	0.43	0.42	0.42
CTPN [50]	0.54	0.51	0.52	0.55	0.53	0.54
TextSnake[9]	0.68	0.85	0.76	0.69	0.85	0.77
TextField[34]	0.83	0.79	0.81	0.83	0.80	0.81
Xue et al.[36]	0.52	0.50	0.51	0.53	0.51	0.52
SWT[61]	0.09	0.21	0.13	-	-	-
CTD+TLOC [62]	0.70	0.77	0.73	-	-	-
PSENet-1s [59]	0.82	0.79	0.80	-	-	-
Proposed Method	0.56	0.51	0.53	0.56	0.52	0.54

method cannot beat TextField and TextSnake methods as these methods cannot handle both kinds of images, properly. Therefore, the proposed method neither scores high nor low compared with the results of existing methods for all the standard datasets. Thus, one can conclude that the proposed method is competitive for text detection in daylight images and the best for low-light images.

E. Discussion

When we analyze the results of the proposed and existing methods on the ICDAR 2013, ICDAR 2015, SVT, ICDAR 2017-MLT, Total-Text and CTW1500 datasets, almost all of the methods score slightly better results after enhancement compared with before enhancement as reported in Table VI Table XI. It is expected that if a dataset involves more low contrast, low resolution and poor quality images, the proposed enhancement helps in improving text detection performance. Therefore, the proposed method obtains consistent results for both before and after enhancement, while the existing methods do not.

Overall, if a dataset contains images with poor quality caused by low resolution, low contrast and distortion, the proposed enhancement helps in improving text detection performance. Our method is good for low-light images and is competitive for daylight images. As has been demonstrated, the proposed method scores consistent results before and after enhancement, especially with respect to F-measure. Therefore, we can conclude

that the proposed enhancement method improves text detection in both low-light and daylight images.

V. Conclusion and Future Work

In this work, we have proposed a new method for detecting arbitrarily-oriented text in low-light images through image enhancement. For enhancement, our method employs a combination of spatial and frequency domain-based features. For detecting text in enhanced images, we use MSER for identifying text candidates. In this work we also introduce the COLD concept in the polar domain for extracting features from text candidates. To improve the performance of text detection, the proposed approach integrates features extracted by COLD and deep features extracted by a fully connected convolution neural network. Experimental results on the output of the enhancement step show that it is better than existing methods in terms of standard quality measures. In the same way, the experiments on text detection in low-light images and standard benchmark natural scene images show that our method is better for the former compared with to existing methods. For natural scene images, our method is consistent both before and after enhancement. However, according to the experimental results, the proposed method falls short of existing methods on benchmark natural scene datasets that do not contain low-light images. This provides the next target in the quest to develop fully robust text detection techniques.

Acknowledgement

The authors would like to thank anonymous reviewers for their constructing comments and suggestions to improve the quality and clarity of the work.

References

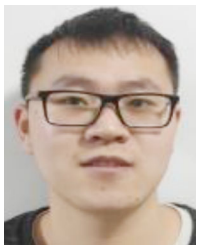
[1] X. Jiang, H. Yao, S. Zhang, X. Lu and W. Zeng, "Night video enhancement using improved dark channel prior," in *Proc. Int. Conf. Image Process.*, 2013, pp. 553–557.
[2] N. Boonsim and S. Prakoonwit, "Car make and model recognition under limited lighting conditions at night," *Pattern Anal. Appl.*, vol. 20, pp. 1195–1207, 2017.
[3] S. Tian, S. Lu and C. Li, "WeText: Scene text detection under weak supervision," in *Proc. Comput. Vis. Pattern Recognit.*, 2017, pp. 1501–1509.
[4] D. Deng, H. Liu, X. Li, and D. Cai, "PixelLink: Detecting scene text via instance segmentation," in *Proc. AAAI*, 2018, pp. 6773–6780.
[5] X. Zhou *et al.*, "East: an efficient and accurate scene text detector," in *Proc. Comput. Vis. Pattern Recognit.*, 2017, pp. 2642–2651.
[6] B. Shi, X. Bai and S. Belongie, "Detecting Oriented Text in Natural Images by Linking Segments", in *Proc. Comput. Vis. Pattern Recognit.*, 2017, pp. 3482–3490.
[7] Y. Liu, L. Jin, S. Zhang, and S. Zhang, "Detecting curve text in the wild: New dataset and new solution," *arXiv:1712.02170*, 2017.
[8] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Multi-oriented and multilingual scene text detection with direct regression," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5406–5419, Nov. 2018.
[9] Y. Xu *et al.*, "TextField: Learning a deep direction field for irregular scene text detection," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5566–5579, 2019.
[10] K. S. Raghunandan *et al.*, "Multi-script-oriented text detection and recognition in video/scene/born digital images," *IEEE Trans. Circuits Sys. Video Technol.*, vol. 29, no. 4, pp. 1145–1162, Apr. 2019.
[11] J. Tang *et al.*, "SegLink: Detecting dense and arbitrary-shaped scene text by instance-aware component grouping," *Pattern Recognit.*, vol. 96, 2019, Art. no. 106954.
[12] M. Liao, Z. Wan, C. Yao, K. Chen, and X. Bai, "Real-time scene text detection with differentiable binarization," in *Proc. AAAI*, 2020, pp. 1–8.
[13] Y. Liu *et al.*, "ABCNet: Real time scene text spotting with adaptive Bezier curve network," in *Proc. Comput. Vis. Pattern Recognit.*, 2020, pp. 9809–9818.
[14] F. Zhan *et al.*, "GA-DAN: Geometry-aware domain adaptation network for scene text detection and recognition," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 9104–9113.
[15] X. Dong *et al.*, "Fast efficient algorithm for enhancement of low lighting video," in *Proc. Int. Conf. Multimedia Expo*, 2011, pp. 1–6.
[16] S. Sharma, J. J. Zuo, and G. Fang, "Contrast enhancement using pixel based image fusion in wavelet domain," in *Proc. Int. Conf. Contemporary Comput. Inform.*, 2016, pp. 285–290.
[17] W. Rui and W. Guoyu, "Medical X-ray image enhancement method based on TV-Homomorphic filter," in *Proc. Int. Conf. Image, Vis. Comput.*, 2017, pp. 315–318.
[18] P. Ravishankar, R. S. Sharmila, and V. Rajendran, "Acoustic image enhancement using Gaussian and Laplacian pyramid-a multiresolution based technique," *Multimedia Tools Appl.*, vol. 77, no. 5, pp. 5547–5561, 2018.
[19] K. S. Raghunandan *et al.*, "Riesz fractional based model for enhancing license plate detection and recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2276–2288, Sep. 2018.
[20] C. Zhang *et al.*, "New fusion based enhancement for text detection in night video footage," in *Proc. Pacific Rim Conf. Multimedia*, 2018, pp. 46–56.
[21] Y. Gao, Y. Chen, J. Wang, M. Tang, and H. Lu, "Reading scene text with fully convolutional sequence modeling," *Neurocomput.*, vol. 339, pp. 161–170, 2019.
[22] L. Deng *et al.*, "Detecting multi-oriented text with corner-based region proposals," *Neurocomput.*, vol. 21, pp. 134–42, 2019.
[23] C. Xue, S. Lu, and W. Zhang, "MSR: Multi-scale regression for scene text detection," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2019, pp. 989–995.
[24] J. Ma *et al.*, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3111–3122, 2018.
[25] D. Bazazian *et al.*, "FAST: Facilitated and accurate scene text proposals through FCN guided pruning," *Pattern Recognit. Lett.*, vol. 119, pp. 112–120, 2019.
[26] Y. Liu, L. Jin, S. Zhang, C. Luo, and S. Zhang, "Curved scene text detection via transverse and longitudinal sequence connection," *Pattern Recognit.*, vol. 90,, pp. 337–345, 2019.
[27] S. Tian *et al.*, "Text Flow: A unified text detection system in natural scene images," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 4651–4659.
[28] C. Xue, S. Li, and F. Zhan, "Accurate scene text detection through border semantics awareness and boostrapping," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 370–387.
[29] D. N. Van, S. Lu, S. Tian, N. Ouarti, and M. Mokhatari, "A pooling based scene text proposal technique for scene text reading in wild," *Pattern Recognit.*, vol. 87, pp. 118–129, 2019.
[30] M. Liao, Z. zhu, B. Shi, G. S. Xia, and X. Bai, "Rotation-sensitive regression for oriented scene text detection," in *Proc. Comput. Vis. Pattern Recognit.*, 2018, pp. 5909–5917.
[31] M. Liao, B. Shi, and X. Bai, "TextBoxes++: A single-shot oriented scene text detector," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3676–3690, Aug. 2018.
[32] P. Lyu, C. Yao, W. Wu, S. Yan, and X. Bai, "Multi-oriented scene text detection via corner localization and region segmentation," in *Proc. Comput. Vis. Pattern Recognit.*, 2018, pp 7553–7562.
[33] P. Lyu, M. Liao, W. Wu, and X. Bai, "Mask TextSpotter: An end-to-end trainable neural network for spotting text with arbitrary shapes," in *Proc. Eur. Conf. Comput. Vis.*, 2018 pp 1–17.
[34] S. Long *et al.*, "Textsnake: A flexible representation for detecting text of arbitrary shapes," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 20–36.
[35] P. Shivakumara *et al.*, "Fractional means based method for multi-oriented keyword spotting in video/scene/license plate images," *Expert Syst. Appl.*, vol. 118, pp. 1–19, 2019.
[36] M. Xue, P. Shivakumara, C. Zhang, T. Lu, and U. Pal, "Curved text detection in blurred/non-blurred video/scene images," *Multimedia Tools Appl.*, vol. 78, pp. 1–25, 2019.
[37] S. Huang, H. Xu, X. Xia, and Y. Zhang, "End-to-end vessel plate number detection and recognition using deep convolutional neural networks and LSTMs," in *Proc. 11th Int. Symp. Comput. Intell. Design*, 2018, pp. 195–199.
[38] R. Panahi and I. Gholampour, "Accurate detection and recognition of dirty vehicle plate numbers for high speed applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 4, pp. 767–779, Apr. 2017.

[39] Wahyono and K. Jo, "LED dot matrix text recognition method in natural scene," *Neurocomput.*, vol. 151, pp. 1033–1041, 2015.
[40] M. S. A. Shemarry, Y. Li, and S. Abdulla, "Ensemble of adaboost cascades of 3L-LBPs classifiers for license plated detection with low quality images," *Expert Syst. Appl.*, vol. 92, pp. 216–235, 2018.
[41] C. H. Lin, Y. S. Lin and W. C. Liu, "An efficient license plate recognition system using convolutional neural networks," in *Proc. Int. Conf. Appl. Syst. Invention*, 2018, pp. 224–227.
[42] S. Mohanty, T. Dutta, and H. P. Gupta, "An efficient system for hazy scene text detection using a deep CNN and patch-NMS," in *Proc. 24th Int. Conf. Pattern Recognit.*, 2018, pp. 2588–2593.
[43] S. He and L. Schomaker, "Beyond OCR: Multifaceted understanding of handwritten document characteristics," *Pattern Recognit.*, vol. 63, pp. 321–333, 2017.
[44] M. Donoser and H. Bischof, "Efficient maximally stable extremal region (MSER) tracking," in *Proc. Comput. Vis. Pattern Recognit.*, 2006, pp. 1–8.
[45] C. Yan *et al.*, "Effective Uyghur language text detection in complex background images for traffic prompt identification," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 220–229, Jan. 2018.
[46] [Online]. Available: https://pan.baidu.com/s/1Or69VMOXPFDCGLjaSAvY9A code:8bk5, Accessed on: 2020.
[47] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.
[48] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a completely blind image quality analyzer," *IEEE Signal Process. Lett.*, vol. 22, no. 3, pp. 209–212, Mar. 2013.
[49] N. Venkatanath, D. Praneeth, Bh. M. Chandrasekhar, S. S. Channappayya, and S. S. Medasani, "Blind image quality evaluation using perception based features," in *Proc. 25th Nat. Conf. Commun.*, 2015, pp. 1–6.
[50] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 56–72.
[51] Q. Wang, Y. Lu, and S. Sun, "Text detection in nature scene images using two-stage nontext filtering," in *Proc. 13th Int. Conf. Document Anal. Recognit.*, 2015, pp. 106–110.
[52] A. Zhu, R. Gao, and S. Uchida, "Could scene context be beneficial for scene text detection?" *Pattern Recognit.*, vol. 58, pp. 204–215, 2016.
[53] S. Lu, T. Chen, S. Tian, J. H. Lim, and C. L. Tan, "Scene text extraction based on edges and support vector regression," *Document Anal. Recognit.*, vol. 18, pp. 125–135, 2015.
[54] Y. Wu, W. Wang, P. Shivakumara, and T. Lu, "A robust symmetry-based method for scene/video text detection through neural network," in *Proc. Int. Conf. Document Anal. Recognit.*, 2017, pp. 1249–1254.
[55] P. He *et al.*, "Single shot text detector with regional attention," in *Pro. Int. Conf. Comput. Vis.*, 2017, pp. 3066–3074.
[56] W. Huang, Y. Qiao, and X. Tang, "Robust scene text detection with convolution neural network induced mser trees," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 497–511.
[57] L. Neumann and J. Matas, "Real-time scene text localization and recognition," in *Proc. Comput. Vis. Pattern Recognit.*, 2012, pp. 3538–3545.
[58] Y. Baek, B. Lee, D. Han, S. Yun and H. Lee, "Character region awareness for text detection," *arXiv:1904.01941*, 2019.
[59] X. Li, W. Wang, W. Hou, R. Z. Liu, T. Lu, and J. Yang, "Shape robust text detection with progressive scale expansion network," *arXiv:1806.02559*, 2018.
[60] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "Textboxes: A fast text detector with a single deep neural network," in *Proc. AAAI*, 2017.
[61] E. Boris, O. Eyal, and W. Yonatan, "Detecting text in natural scenes with stroke width transform," in *Proc. Comput. Vis. Pattern Recognit.*, 2010. pp. 2963–2970.
[62] X. Liu *et al.*, "FOTS: Fast oriented text spotting with a unified network," *arXiv:1801.01671*, 2018.

Minglong Xue is the Ph.D. candidate with Department of Computer Science and Technology, Nanjing University, China. His area of interest includes image processing, pattern recognition and video text understanding.

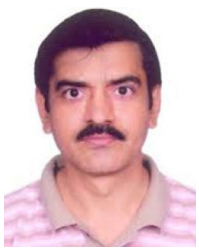
Palaiahnakote Shivakumara received the B.Sc., M.Sc., M.Sc. Technology, and Ph.D. degrees in computer science from the University of Mysore, Mysore, India, in 1995, 1999, 2001, and 2005, respectively. He is currently an Associate Professor with the Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia. Previously, he was with the Department of Computer Science, School of Computing, National University of Singapore from 2008 to 2013 as a Research Fellow on Video text extraction and recognition project. He has been an 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 authored or coauthored 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.

Chao Zhang received the postgraduate degree from the Department of Computer Science and Technology, Nanjing University. His research interests include image processing and pattern recognition.



Yao Xiao is a research student with Nanjing University. His area of interest includes pattern recognition, machine learning, and feature selection.

Tong Lu received the B.Sc. and Ph.D. degrees in computer science from Nanjing University, Nanjing, China, in 1997 and 2005, respectively. He served as an Associate Professor and an Assistant Professor with the Department of Computer Science and Technology, Nanjing University from 2007 to 2005. He is currently a Full Professor with Nanjing University. He also has served as a Visiting Scholar with 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 authored or coauthored more than 130 papers and authored two 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 (Senior Member, IEEE) received the Ph.D. degree from Indian Statistical Institute. He did his Postdoctoral with INRIA (Institut National de Recherche en Informatiqueet en Automatique), France. From January 1997, he is a Faculty member of the Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata and he is currently a Professor and the Head. 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, such as PR, *Electronic Letters on Computer Vision and Image Analysis;* IJDAR, PRL, ACM *Transactions on Asian Language Information Processing,* etc. He is a fellow of IAPR.

Daniel Lopresti received the Ph.D. degree in computer science from Princeton University, in 1987. After completing his doctorate, he joined the faculty with Brown University and taught courses ranging from VLSI design to computational aspects of molecular biology and conducted research in parallel computing and VLSI CAD. He went on to help found the Matsushita Information Technology Laboratory in Princeton, and later also served on the research staff at Bell Labs where his work turned to document analysis, handwriting recognition, and biometric security. In 2003, Dr. Lopresti joined the Department of Computer Science and Engineering with Lehigh University, where his research examines fundamental algorithmic and systems-related questions in pattern recognition, bioinformatics, and security. He is Co-Editor-in-Chief of the International Journal on Document Analysis and Recognition. He has applied his technical expertise on the controversial topic of electronic voting, and has recently been working to help develop the Code 8.7 international collaboration to apply AI in the fight against human trafficking. Since 2015, he has served on the Computing Community Consortium Council of the Computing Research Association and is currently a member of the CCC Executive Committee. He is also on the Executive Committee of the International Association of Pattern Recognition and currently serves as IAPR's Treasurer.

Zhibo Yang received the B.S. degree from the Department of Automation, Harbin Institute of Technology, Harbin, China, in 2010, and M.S. degree from the Department of Automation, Beijing, China, in 2014. He is currently a Senior Algorithm Engineer in Machine Intelligence Technology Department, DAMO, Alibaba Group, Hangzhou, China. His research interests include object detection and text detection in images/videos.