Local Gradient Difference Features for Classification of 2D-3D Natural Scene Text Images

[1]Lokesh NANDANWAR, [1]Palaiahnakote Shivakumara, [2]Ramachandra Raghavendra, [3]Tong Lu, [4]Umapada Pal and [5]Daniel Lopresti and [1]Nor Badrul Anuar

[1]Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia. Email: lokeshnandanwar150@gmail.com, shiva@um.edu.my, badrul@um.edu.my.

[2]Faculty of Information Technology and Electrical Engineering, IIK, NTNU, Norway, raghavendra.ramachandra@ntnu.no

[3]National Key Lab for Novel Software Technology, Nanjing University, Nanjing, China. Email: lutong@nju.edu.cn

[4]Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India. Email: umapada@isical.ac.in

[5]Computer Science & Engineering, Lehigh University, Bethlehem, PA, USA. Email: lopresti@cse.lehigh.edu.

***Abstract*—Methods developed for normal 2D text detection do not work well for text that is rendered using decorative, 3D effects, etc. This paper proposes a new method for classification of 2D and 3D natural scene text images so that an appropriate recognition method can be chosen accordingly based on the classification results for better performance. The proposed method explores local gradient differences for obtaining candidate pixels, which represent a stroke. To study the spatial distribution of candidate pixels, we propose a measure, called COLD, which is denser for pixels toward the center of strokes and scattered for non-stroke pixels. This observation leads us to introduce mass features for extracting the regular spatial pattern of COLD, which indicates a 2D text image. The extracted features are fed into a Neural Network (NN) for classification. The proposed method is tested on (i) a new dataset introduced in this work (ii) a second dataset assembled from standard natural scene datasets (iii) Non-Text Image datasets which does not contain text, rather it contains objects. Experimental results of the proposed method on images with text and non-text show that the proposed method is independent of text. The proposed approach improves text detection and recognition performance significantly after classification.**

Keywords—Gradient, edges, Strokes Cold, Text detection, 2D-3D image classification

I. INTRODUCTION

Text detection and recognition in natural scene images and videos has received substantial attention from researchers because of the potential for real-world applications, such as navigation for autonomous vehicles and assisting tourists in understanding local-language signage via Optical Character Recognition (OCR) on their mobile phones. The field has matured and there are powerful methods for addressing many of its challenges. For example, the methods proposed in [1-7] address the challenges of arbitrarily-shaped text lines, text detection from multiple views, bib number detection in marathon images, text detection in blurred and non-blurred scene images, text detection and recognition in web videos, and text detection in images affected by perspective distortion.

One can understand from the above methods that the approaches work well for the images having 2D text. But in reality, it is common that images can contain 3D text due to 3D movies and 3D video. If the image contains text with shadow and effect of depth information, it is considered as 3D image else it is considered as 2D image. The methods proposed for 2D text images may not perform well for 3D text images due to the presence of shadow and decorative characters of the 3D effect. Therefore, there is a dearth for text detection in 3D video and naturals scene images. The reason is that due to the presence of shadows and depth information in the image, we can expect a loss of character shapes. It is evident from the sample results shown in Fig. 1, where it can be seen that the text detection method called PSENet [1], works well for the 2D image while it does not detect text in the 3D image as shown in Fig. 1(a).

(a) Text detection by PSEnet method [1] in 2D and 3D text images before classification.

(b) Text detection by PSEnet method [1] in 2D and 3D text images after classification.

Fig. 1. Text detection performance before and after classification of 2D and 3D images

Therefore, we modify the same method [1] to train on individual classes of 2D and 3D images for detecting text in 2D and 3D images as shown in Fig. 1(b), where it is noted that the modified method performs well for both 2D and 3D images. It has motivated us to propose a new classification method for classifying 2D and 3D images such that text detection and recognition performance improves for both 2D and 3D images. In the same way, one can think of developing one separate method for 3D text detection by taking the advantage of 2D and 3D image classification to achieve better results instead of modifying existing methods. Since developing a universal method for handling challenges of both 2D and 3D images is hard, this work proposes the classification of 2D and 3D images.

II. RELATED WORK

As discussed in the introduction, most existing methods focus on 2D text images. Liu et al. [8] proposed a unified network for

text spotting in natural scene images. It aims at combining text detection and recognition by designing a single network for achieving better results. Li et al. [1] introduced a progressive scale expansion network for robust text detection in natural scene images. The method focuses on addressing the challenge of fixing bounding boxes for arbitrarily shaped text. Xu et al. [9] proposed a deep directional field for irregular text detection in natural scene images. To find a solution to arbitrary orientation and shaped text, the method explores direction information through a deep network. Raghunandan et al. [10] proposed a bit plane based method for text detection in a natural scene, video, and born-digital images. The method uses convex and concave deficiencies for identifying candidate bit plane and then the same plane is used for text detection. Villamizar et al. [11] explored a multi-scale sequential network for text detection in natural scene images.

The method works based on classifying text into different classes according to semantics. To achieve this, the method explores the deep learning model. Roy et al. [2] proposed a method for text detection in multi-view of scenes by exploring Delaunay triangulation and its characteristics. The goal of this method is to address the challenges of the same text in different views. It extracts features using Delaunay triangulation and uses similarity measures for achieving the results. Wang et al. [12] use the two-state network for text detection in natural scene images. It fixes a quadrilateral boundary for the text to solve the issue of arbitrary shape text. Liao et al. [13] proposed differentiable binarization for text detection in natural scene images. It fixes the automatic threshold for binarization to improve text detection performance.

From the above review, it is noted that though the methods explored deep learning models in a different way for addressing challenges of text detection. The approaches do not consider 3D images for text detection. Therefore, there are two ways of finding a solution to the text detection in 3D images. (i) Classifying the 2D and 3D images such that an appropriate method can be used to improve text detection performance. (ii) Developing a unified or universal method for detecting text in both 2D and 3D images. In this work, we choose former one because developing a generalized method is not advisable, at the same time, we can make use of existing methods by modifying to detect text in 3D images. Hence, this work aims at proposing a new method for the classification of 2D and 3D images.

However, there is an attempt to classify 2D and 3D texts. Xu et al. [14] where a method for multi-oriented graphics-scene 3D text classification in the video is proposed. The method explores the medial axis points of a character for classifying 2D and 3D texts in the video. Zhong et al. [15] proposed to use shadow detection for 3D text classification. The approach fixes some thresholds for detecting shadow pixels to estimate the depth of the 3D text information. However, the scope of the above method is limited to detected text but not images. With the same objective, Nandanwar et al. [16] proposed a method for 2D and 3D text classification based on common point detection. The method explores Gradient Vector Flow (GVF) to find common points for text and its shadow, and then statistical features of common points are used for classification. This work is confined to detected text but not full images containing 2D and 3D texts as our work. On the other hand, since the objective of the above methods is to improve text recognition performance through classification, which is same as the proposed work, inspired us to propose a method for classifying images containing 2D and 3D images in this work.

III. The Proposed Method

For the image containing 2D text, one can expect regularity in stroke width distances because it is true that the stroke width of characters in the text is almost same [16], while for the image containing 3D text, it does not due to the presence of shadow, decorative characters using depth information. This observation motivated us to detect candidate pixels in the images and distances between the candidate pixels. Inspired by the work [17], where the local resultant force has been introduced to detect the boundary point of the objects in the medical images, we explore the same concept by considering gradient difference rather than direction information for detecting candidate points in the images. Since the stroke width of the characters in the 2D text shares the same spatial proximity, it is expected dense cluster at center in polar domain, while scattered clusters for 3D text in the images.

Motivated by the method called COLD (cloud of line distribution), [18] which was developed for the problem of writer identification using handwriting and is being adapted here to the classification of 2D and 3D text images. To extract the above observation, we propose to extract mass features [19] from the cold distribution. The reason to extract the mass features is that the probability of the pixels which share the same distances is higher than the pixels which have arbitrary distances in the defined granularity. The granularity is defined by the radius of estimated automatically based on distribution of the pixels in the COLD. The extracted mass feature from the granules defined by the radius values fed to Neural Network classifier for classification of 2D and 3D text images. The main advantage of this work is that it works well for the image without text information. The distribution of pixels in the COLD does not make much difference between 2D text and 2D objects in the images.

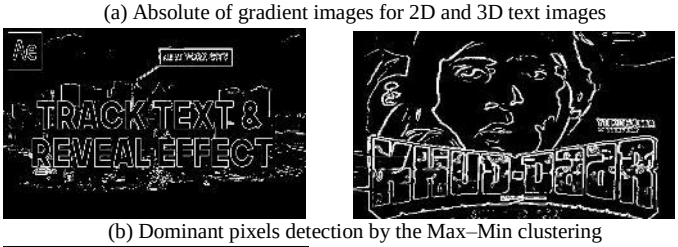
A. Local Gradient Difference for Candidate Pixel Detection

For the input images shown in Fig. 1(a), the proposed method obtains an absolute gradient image for 2D and 3D images as shown in Fig. 2, where it can be seen the pixels which represent edges are sharpen compared to the background pixels. Since the step, namely, Local Gradient Difference (LGD) requires the dominant pixel which represents edges in the images, the proposed method performs Max-Min clustering on absolute gradient images to detecting dominant pixels. For each pixel in the absolute gradient images, we define a 3×3 window and then the proposed method chooses Max value and Min values from the window. The values in the window are compared to with the Max and Min values. If the value is closest to Max value, it is classified into Max cluster else Min cluster. The pixels which classified into Max cluster are considered as dominant pixels as shown in Fig. 2(b) where we can see this process outputs almost all edge pixels for both 2D and 3D text images.

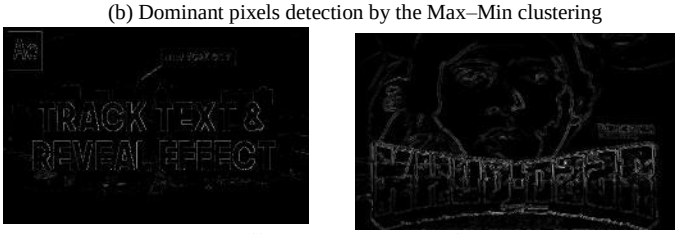
Again for the 3×3 window of each dominant pixels (white pixels), the proposed method computes LGD as defined in

Equation (1), where the gradient difference is calculated using neighboring pixels values. Note that we consider gradient values in the absolute gradient images corresponding to dominant pixels for computing LGD. The effect of LGD can be seen in Fig. 2(c), where it can be noted that dominant pixels preserve the sharpness. With LGD values, the proposed method finds a change in local gradient difference which is called as Local Gradient Resultant (LGR) as defined in Equation (2). The results in Fig. 2(d) show that the LGR process increases the sharpness for the dominant pixels compared to the LGD results. Therefore, one can conclude that the LGR process widens the gap between edge pixels and background pixels in the images. In order to extract such edges pixels, the proposed method employs k-means clustering with k=2 over LGR images, which outputs two clusters. The cluster which gives high mean is considered cluster with candidate pixels as shown in Fig. 2(e) where we can see all edge pixels are classified as candidate pixels.

(a) Absolute of gradient images for 2D and 3D text images

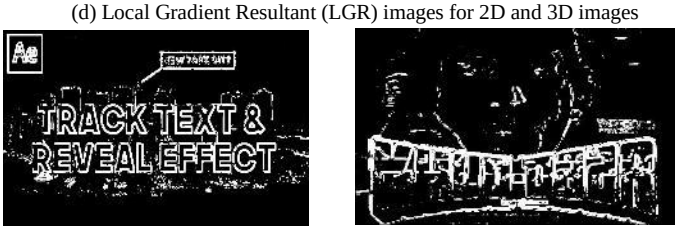


(b) Dominant pixels detection by the Max–Min clustering



(c) Local Gradient Difference (LGD) images for 2D and 3D images

(d) Local Gradient Resultant (LGR) images for 2D and 3D images



(e) Max cluster results given by K-means clustering on LGR images.

Fig. 2. Candidate pixels detection based on local gradient difference.

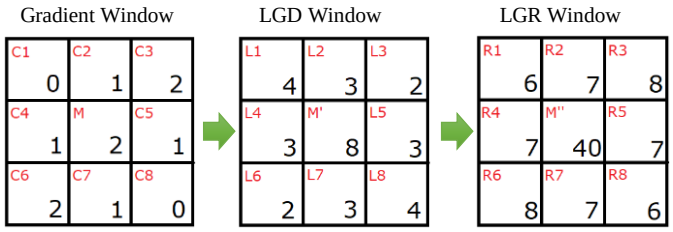
When we compare the results in Fig. 2(e) with the results in Fig. 2(b), unwanted background pixels are removed while edge pixels which represent boundary of the characters preserved.

$$LGD(x,y) = \sum_{i=x-1}^{x+1}\sum_{j=y-1}^{y+1}\{|G(i,j) - G(x,y)|\} \quad \forall\,(x,y) \leftrightarrow MM(x,y) = 1 \quad (1)$$

Where G(x, y) denote absolute gradient image and MM(x, y) denotes Max cluster image.

$$LGR(x,y) = \sum_{i=x-1}^{x+1}\sum_{j=y-1}^{y+1}|LGD(i,j) - LGD(x,y)| \quad (2)$$

The steps for calculating LGD and LGR are shown in Fig. 3 where then gradient difference is computed using neighboring gradient values. For instance, the value of 4 in the first cell of the LGD window is calculated by taking the absolute difference between |0-1| + |0-2| + |0-1| in gradient window. The value of the 0 in the first cell of the gradient window is considered as a center pixel to be replaced with the new value. Similarly, if we consider the value of 6 as a center pixel in the LGR window, it can be calculated by taking the absolute difference between |4-3| + |4-8| + |4-3| in gradient window. If we consider the value of 8 as center pixel in LGD, the LGD for this value is can calculated as defined equations in Fig. 3, which is the absolute difference between |2-0| + |2-1| + |2-2] + |2-1 + |2-0| + |2-1| + |2-2] + |2-1| in gradient window. For calculating LGD values, the proposed method uses gradient values and for LGR values, the LGD values are used. It is expected the above steps should output the boundary points of characters and the objects in the images. Fig. 2(e) shows that most of the boundary points are detected irrespective of character and objects in the images. Therefore, one can argue that these steps work well for the images containing any object and text.



$$LGD(M) = M' = |M - C1| + |M - C2| + |M - C3| + |M - C4| + |M - C5| + |M - C6| + |M - C7| + |M - C8| = 8$$

$$LGR(M') = M'' = |M' - L1| + |M' - L2| + |M' - L3| + |M' - L4| + |M' - L5| + |M' - L6| + |M' - L7| + |M' - L8| = 40$$

Fig. 3. Illustration for LGD and LGR for 3×3 Gradient window

B. COLD for Extracting Spatial Proximiy of Candidate Pixels

When we look at the candidate pixels arrangements in Fig. 2(e), it is clear that the spatial proximity between the candidate pixels in the 2D image is close compared to the candidate pixels in 3D images. This is the cue for differentiating 2D and 3D images, which is the same for the images of any objects including text. To extract such observation, we propose to explore the concept of COLD (cloud of line distribution). COLD uses the distance between the candidate pixels and spatial coordinates to obtain a distribution in the polar domain as defined in Equation (3) and

Equation (4), where θ is the angle between two pixels and r is the distance between two pixels.

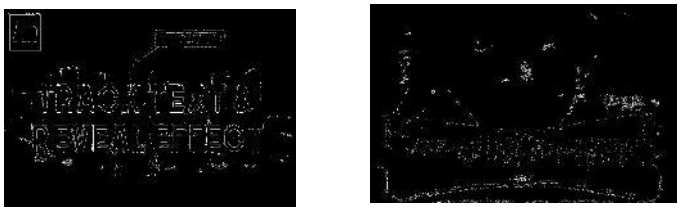
$$\theta = \tan^{-1}\left(\frac{y_{i+1}-y_i}{x_{i+1}-x_i}\right) \quad (3)$$

$$r = abs\left(\sqrt{(y_{i+1}-y_i)^2+(x_{i+1}-x_i)^2}\right) \quad (4)$$

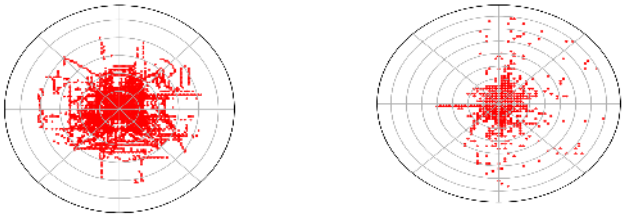
Here $(x_i, y_i)$and (x_{i+1}, y_{i+1}) denote the coordinates of a pair pixels. When we draw points for all the pairs in polar domain (θ, r) it results in a distribution.

(a) Connected component labeling for candidate pixels images

(b) Traversing in 360 directions to find stroke pixels for each connected component. Yellow color denotes direction and red pixel denote stroke pixels.



(c) Stroke pixels for the 2D and 3D images



(d) Cold distribution for the stroke pixels of 2D and 3D images.

Fig. 4. Studying spatial distribution of stroke pixels through COLD.

Since the distance between all the candidate pixels does not contribute to classification, the proposed method focuses on extracting the distance which represents the stroke width of characters and objects in the images. The proposed method considers each component in a candidate pixel image as a connected component as shown in Fig. 4(a) where we can see boundaries for each component in an image. Since small connected components do not contribute much for classification, we remove components that are smaller than five pixels. Our method finds centroid for each component by considering x and y coordinates of all the white pixels in the connected component. From each centroid of each component in the image, the proposed method traverses in 360 directions until it reaches the last boundary pixels of the connected component, which is considered as end stroke width points. From the end stroke width point, the proposed method retraces in reverse direction until it finds the transition either 1 to 0 or 0 to 1, which is considered as begin stroke width point. This process is illustrated in Fig. 4(b), where it can be seen that directions from the centroid and the points which represents stroke width pair marked in red color for each direction. It is observed from Fig. 4(b) that the character "R" in the 2D image is considered as one connected component while the whole text line in the 3D image as one connected component. This is due to the presence of shadow and background pixels in 3D image. This reflects in the spatial relationship between the pixels and makes a difference between 2D and 3D images.

The distance between the beginning and end stroke width points are considered as stroke width distance. The points of stroke width pairs are shown in Fig. 4(c) for 2D and 3D images, where we can observe the spatial distribution of pixels are different for 2D and 3D images. We draw points using angle and distances in the polar coordinate system, which results in COLD distribution as shown in Fig. 4(d), where one can see dense clusters at the center for 2D images and scattered clusters for 3D images. The proposed method extracts such observation by estimating mass features for each ring of radius, which will be discussed in the subsequent section.

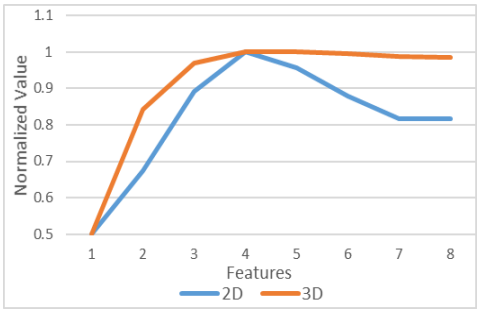


Fig. 5. Mass features for 2D and 3D images.

C. Mass Features Extracted from COLD for Classifiication

After getting the COLD plot, we use it to calculate the Mass features using splits. To decide the radius of rings in COLD we use the mean distance of stroke pixels pairs in the image. The optimalnumber of rings is eight which is determined empirically based on the datasets we used. As defined in [19], $mass(x_a)$ for a ring a, where $x_a \in \{x_1, x_2, \ldots, x_{n-1}, x_n\}$ is defined as a summation of a series of mass base weighted by $p(a)$ over n rings, here $\{x_1, x_2, \ldots, x_{n-1}, x_n\}$ is the number of pixels in rings from range$(1, n)$, where n is equal to 8. The mass is defined as follows:

$$mass(x_a) = \sum_{k=1}^{a-1}(n-a)\times p(a) + \sum_{k=a}^{n} a \times p(a), a \in (1,n) \quad (5)$$

Where $p(a)$ is the probability of number of pixels in ring a

$$p(a) = \frac{x_a}{\sum_{i=1}^{n} x_i} \quad a \in (1,n)$$

Finally, we combine all mass features for all rings $a\,\{a \in (1,n)\}$ to obtain the feature vector for the input image. The line graph of the normalized feature vector for 2D and 3D image is shown in Fig.5, where it can be seen that the mass features have the ability to distinguish 2D and 3D images.

For the classification of 2D and 3D texts, we use a Neural Network by passing extracted features into the network. The structure of the proposed network is in Fig. 6. It has 6hidden dense layers. The Input layer has 8 features and the final output layer has 1 feature {0 for 2D and 1 for 3D}. We use the Rectified linear activation function "ReLU" for all hidden layers and Sigmoid [20] activation function for the final layer. Dropout with a drop rate of 10% in between the layers is used to reduce overfitting. Binary cross-entropy loss(L) is estimated [21] as defined in Equation (6), where Pr(y) is the probability predicted and 'y' is the label for a total number of 'n' samples. Adam [22] optimizer with a learning rate of 0.005 is used during the training process and the batch size used is 12 during training. Each model is trained for a maximum of 100 epochs with a model checkpoint to store the best-trained model using Keras framework. For training we used 80% of the samples chosen randomly, and for testing the remaining20% of the samples. The details of the proposed neural network are shown in Fig. 6.

$$L(q) = -\frac{1}{n}\sum_{k=1}^{n}(y_k \,.\log(Pr(y_k)) + (1-y_k).\log(1-Pr(y_k))) \qquad (6)$$

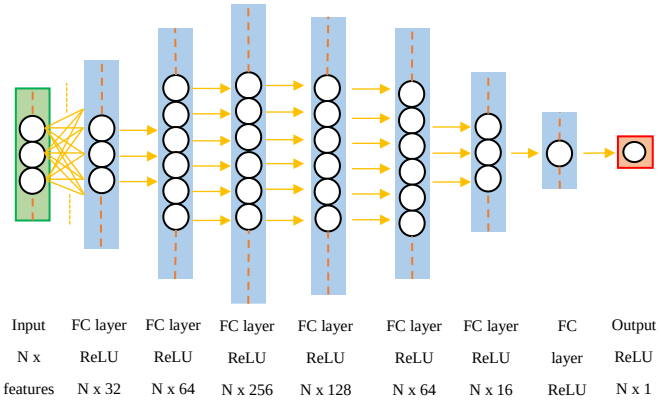


Fig. 6. Architecture of Neural Network Classifier

IV. Experimental Results

Since the problem of 3D text image classification from 2D text image is new, we collect our dataset from different resources, such as YouTube, movie posters, and the internet. To choose 3D images, we check whether the image exhibits shadows and depth. If an image contains such an effect, it is considered as 3D. Otherwise it is considered as 2D.The collected images include complex backgrounds, low resolution, low contrast, font, font-size variations, arbitrarily shaped characters. and some extent to distortion. To test the robustness of the proposed method, we also choose 3D images from the benchmark datasets of natural scene images, namely, ICDAR 2013, ICDAR 2015, ICDAR 2017 MLT, and COCO-Text if the dataset contains any 3D images. For 2D image collection, we choose randomly from the same datasets. This dataset is considered as a standard dataset for experimentation at the image level.

At the same time, to evaluate the proposed method at the text line level, the detected text lines from our dataset and the same aforementioned natural scene datasets are considered for experimentation. In addition, we also use the text lines that are used in [15] for 2D and 3D text line classification. This dataset contains examples of text with varying shadows, font sizes and backgrounds. This dataset is used for classifying text line images through shadow detection to improve text recognition. Similarly, to show the proposed method does not depend ont ext for the classification of 2D and 3D images, we choose 2D and 3D images without text information from [24] where natural scene images are used for categorizing 15 kinds of scenes This dataset consists of images with 2D and 3D office, kitchen, living room, mountain, etc. Sample images for each dataset including our dataset and images of non-text information are shown in Fig. 7(a)-Fig.7(b), where we can see all the images exhibit different complexities in terms of foreground (text or object) and background variations.

Fig. 7. Sample images of our, natural scene and non-text images datasets. All the sample images are classified successfully by the proposed

In summary, the proposed work comprises 1056 text images, 3456 text line images, and 500 images without text information (non-text images) for evaluation. Details of the sizes of 2D and 3D classes for each dataset can be seen in Table I. With this information, we can infer that the considered dataset is fair enough for evaluating the proposed and existing methods for the classification of 2D and 3D images of different types at different levels.

Table I. Details of different types of dataset for evaluation.

Datasets	Type	2D	3D	Total
Our dataset	Image	400	400	800
Standard Natural Scene	Image	130	126	256
Our	Line	513	505	1018
IIIT5k	Line	317	305	632
COCO-Text	Line	472	530	992
ICDAR 2013	Line	123	74	197
ICDAR 2015	Line	111	90	201
Zhong et al. [15]	Line	200	216	416
Ali et al. [24]	Non-Text Images	250	250	500

We implemented the following existing methods for comparative study with the proposed method to show that the proposed method is effective. Xu et al. [14] that explores gradient direction for classifying 2D and 3D images having text information. Zhong et al. [15] that explores shadow detection information for 2D and 3D image classification. The motivation to choose the above two methods is that the objective of the methods is the same as the proposed method. That is the classification of 2D and 3D images for improving text detection and recognition performance. The same two existing methods are used for classifying images at the text line level also. In the case of the text line level, we run the text detection method [1] to obtain text lines from 2D input images and we segment text lines manually from 3D images. The segmented text lines are input for classification.

To show the usefulness of the proposed classification at the image level, we run the following text detection methods before and after classification. For the before classification experiments, the methods consider both 2D and 3D images together as the input for classification, while for after classification experiments, images of individual classes are considered as the input for text detection. Note that for before classification experiments, the text detection methods are trained on both 2D and 3D images whilst for after classification, the methods are trained on images of respective classes. Wang et al. [1], Liu et al. [8], and Liao et al. [13] explore deep learning for text detection in natural scene images. The reason to choose the above-mentioned methods is that implementations are publicly available and these are the state-of-the-art methods. It is expected that the text detection methods should report worse results before classification and better results after classification. This is expected because the complexity of the classification problem increases when we consider both 2D and 3D images together as the input for text detection, whereas the complexity is reduced when we classify images as 2D and 3D.In the same way, to show the usefulness of the proposed classification, we run two recognition methods, namely, ASTER [6] and MORAN [7], which employ deep learning models for text of different complexities. These methods are robust to complex backgrounds, low resolution, low contrast, arbitrarily shaped characters and orientations similar to our proposed approach. In addition, the implementations software would be made available to the public. Since the dataset is small from the standpoint of deep learning, we use augmentation techniques to increase the number of samples by generating synthetic images with different operations, such as rotation, scaling, and shear transform.

For evaluating the performance of the proposed and existing methods for classification, we usethe well-known measures, namely, confusion matrix and average classification rate. The classification rate is defined as the number of images classified correctly by the proposed method divided by the actual number of images. The average classification rate is defined as the mean of diagonal elements of the confusion matrix. For text detection experiments, we use standard F-measures, which is the harmonic mean of recall and precision for evaluation. For recognition experiments, we use the recognition rate, which is defined as the number of characters recognized correctly divided by the actual number of characters. For training the proposed method as described in the Proposed Methodology section, the following values are determined. For the text detection experiments, each model is trained for 10 epochs with an initial learning rate of 0.0005 and a batch size of 8. For the text recognition experiments, each model is trained for 15 epochs with 0.01 as an initial learning rate and with 16 as the batch size. While training, we use checkpoints to store the best performing model for later use during testing.

A. Ablation Study

To analyze the contribution and effect of each step involved in the proposed method, we conduct the following experiments on our dataset. In the proposed work, Local Gradient Difference (LGD), Cloud of Line Distribution (COLD) and Mass features are the key steps. To assess the contribution of each key step, generate a confusion matrix, and calculate theaverage classification rate using our dataset as reported in Table II. For experiments without the LGD step, the proposed method considers normalized gradient images as input for the K-means clustering to detect candidate pixels, and then the same steps are used for classification. For experiments without COLD, the proposed method considers stroke pixels given by directions as input for Mass feature extraction and then classification. For experiments with density, the proposed method counts the number of pixels in each ring instead of mass estimation for classification. The results for each key steps are reported in Table II where we can confirm from the average classification rate that each key step contributes to achieving high results (81.875%) in terms of average classification rate.However, it is also noted that the key steps alone are not enough to achieve better results as the proposed method.

Table II. Confusion matrix and average classification rate for the key steps of the proposed method on our dataset at image level(in %).

Classes	Proposed without LGD		Proposed without COLD		Proposed with density		Proposed Method	
	2D	3D	2D	3D	2D	3D	2D	3D
2D	**61.25**	38.75	**76.25**	23.75	**77.5**	22.5	**85.0**	15.0
3D	47.5	**52.5**	41.25	**58.75**	27.5	**72.5**	21.25	**78.75**
Average	56.875		67.5		75.0		**81.875**	

We also conducted a few more experiments to assess the impact of the neural network classifier used in the proposed method. In order to know the effect of the proposed features, the extracted features are fed to a state-of-the-art deep convolution neural network, namely, MobileNetV2 [23] with Adam optimizer, learning rate = 0.01, and batch size = 8, which results in Experiment-1. Similarly, to analyze the effect of the proposed Neural Network classifier, the extracted features are fed to a

conventional SVM classifier for classification, which results in Experiment-2. To test the contribution of the Adam optimizer used in the proposed classifier, the experiments are conducted by replacing Adam optimizer with SGD [25], which results in Experiment-3. In the same way, we also examine 5 fold-cross-validation, which results in Experiment-4. Classification results along with average classification rates of all the four different experiments are reported in Table III. It is observed from Table III that the experiments just described show poor results compared to the proposed method. This demonstrates that the steps and the components used in the proposed method are effective and contribute significantly to classification.

Table III. Experiments for Ablation study for the proposed method using our dataset (in %).

Experiments	Exp-1		Exp-2		Exp-3		Exp-4		Proposed	
Classes	2D	3D	2D	3D	2D	3D	2D	3D	2D	3D
2D	78.2	21.75	64.5	35.5	80.4	19.6	81.2	18.7	85.0	15.0
3D	28.7	71.25	39.0	61.0	24.2	75.8	22.5	77.5	21.2	78.7
Average	74.7		62.7		78.1		79.3		**81.8**	

B. Experiments on Classification

Quantitative results of the proposed and existing methods for our and standard datasetsat image level are reported in Table IV. It is noted from Table IV that for both the datasets, the proposed method achieves the best at average classification rate compared to existing methods. However, the proposed method scores low results for the standard dataset compared to our datasets. This is because the images chosen from the standard dataset are more diversified compared to the images of our dataset. The same is true for existing methods also as these methods report low results for standard dataset. When we compare the results of Zhong et al. and Xu et al. the Zhong et al. is better for our dataset and poor for the standard dataset compared the Xu et al. This is due to the images of standard dataset do not have shadow information while images of our dataset have shadow information. This is valid because Zhong et al. works based on shadow information in the images. On the other hand, since Xu et al. works based on gradient information, it is sensitive to the complex background. Therefore, Xu et al, reports poor results for standard dataset compared to our dataset.

To evaluate the method at the text line level, we generate confusion matrices and calculate the average classification rate for our dataset, as well as the standard and Zhong et al. datasets as reported in Table V and Table VI, respectively. Our method achieves the best average classification rate for all three datasets compared to the existing methods. As expected, the proposed method scores better results at the text line level as opposed to the image level. However, when we look at the outcomes of the existing methods at the image and text line levels, the results are not consistent. This is due to the thresholds and inherent limitations of the methods. On the other hand, in case of the proposed method, since the features are robust to different types of images, it performs well at both the image and text line levels. It is noted from Table VI that Zhong et al.'s method reports poor results for their dataset. This is due to adjusting the many thresholds used in the method. Since their source code is not available, we implemented the method and used the same procedure described earlier to set the thresholds for the comparative study we present in this work.

Table IV. Confusion matrix and average classification rate of the proposed and existing methods on our and standard datasets at image level(in %).

Dataset	Our dataset-image level						Standard dataset-image level					
Methods	Zhong et al.		Xu et al.		Proposed		Zhong et al.		Xu et al.		Proposed	
Classes	2D	3D	2D	3D	2D	3D	2D	3D	2D	3D	2D	3D
2D	76.5	23.5	58.7	41.3	83.0	17.0	52.0	48.0	56.4	33.6	78.4	21.6
3D	41.5	58.5	31.9	68.1	22.0	78.0	42.4	57.6	39.2	60.8	26.5	72.5
Average	67.5		63.4		**80.5**		54.8		58.6		**75.45**	

Table V. Confusion matrix and average classification rate of the proposed and existing methods on our and standard datasets at line level(in %).

Dataset	Our dataset –Text line level						Standard dataset-Text line level					
Methods	Zhong et al.		Xu et al.		Proposed		Zhong et al.		Xu et al.		Proposed	
Classes	2D	3D	2D	3D	2D	3D	2D	3D	2D	3D	2D	3D
2D	32.7	67.2	66.2	33.7	91.1	8.9	53.8	46.1	46.6	53.4	87.7	12.3
3D	46.3	53.6	28.3	71.6	11.2	88.8	32.8	67.2	31.2	68.8	16.5	83.5
Average	43.2		70.48		**89.95**		60.53		57.7		**85.6**	

Table VI. Confusion matrix and average classification rate of the proposed and existing methods on Zhong et al., and Ali et al, Non-Text Image datasets(in %).

Dataset	Zhong et al dataset [15]						Ali et al., Non-Text Image dataset [24]	
Methods	Zhong et al.		Xu et al.		Proposed		Proposed Method	
Classes	2D	3D	2D	3D	2D	3D	2D	3D
2D	67.0	23.0	74.8	25.2	92.9	7.1	82.0	18.0
3D	28.4	65.6	26.0	64.0	12.8	87.2	24.0	76.0
Average	66.3		69.4		**90.05**		78.0	

Since the proposed method does not depend on the text in an image when making the 2D/3D classification, we also tested it on images without text. For this experiment, we use the images that are used in [24] where 2D and 3D non-text images are used for categorizing 15 scenes, as reported in Table VI. The results on Non-Text Images show the proposed method is able to achieve promising results compared to the result on text images. Therefore, we can conclude that our method classifies 2D and 3D images of different types, and hence it is a text independent method. However, the results of non-text images are lower than the results of text images. This is because when the image contains a very little effect of shadow and depth information due to large variations of object shapes, the performance degrades. This is not a major setback of the proposed work because the primary goal of the proposed work is to classify images with text information. Note that the existing methods are not used for comparative study with the proposed method on this dataset because these methods require an image with text or text lines but not the images without text.

C. Text Detection and Recognition Experiments for Validating The Proposed Classification

As discussed earlier, to show the usefulness of the proposed classification method, we perform text detection and recognition experiments before and after classification. For text detection experiments, the full images are input while for text recognition experiments, text lines are input. Quantitative results of different text detection and recognition methods for our and standard datasets are reported in Table VII and Table VIII, respectively. It is observed from Table VII and Table VIII that the performance of text detection and recognition methods improves significantly after classification compared to before

classification for both the datasets. This shows that the proposed classification contributes to improving text detection and recognition performance for 2D and 3D images. At the same time, we can also conclude that single methods are not feasible for achieving better results for 2D and 3D images.

Table VII. Text detection performance in terms of F-measure of different methods for our and standard full dataset at image level before and after classification. BC denotes before classification and AC denotes after classification.

Methods	Our Dataset-image level				Standard dataset-image level			
	BC	AC			BC	AC		
	2D + 3D	2D	3D	Avg	2D + 3D	2D	3D	Avg
PSEnet [1]	67.9	73.3	66.9	70.1	73.3	88.6	64.0	76.3
FOTS [8]	59.2	70.3	56.9	63.6	64.3	75.1	60.5	67.8
DB [13]	60.5	68.2	59.1	63.6	66.6	80.8	61.4	71.1

Table VIII. Text recognition performance in terms of character recognition rate of different methods for our and standard dataset at line levels before and after classification. BC denotes before classification and AC denotes after classification.

Methods	Our Dataset-Text line level				Standard dataset-Text line level			
	BC	AC			BC	AC		
	2D + 3D	2D	3D	Avg	2D + 3D	2D	3D	Avg
ASTER [6]	79.0	97.0	85.7	91.3	88.5	96.1	85.4	90.7
MORAN[7]	87.2	94.1	87.6	90.8	89.7	96.0	86.4	91.2

ACKNOWLEDGMENT

V. CONCLUSION AND FUTURE WORK

In this work, we have proposed a new method for the classification of 2D and 3D text in natural scene images. The proposed method employs a local gradient difference for detecting candidate pixels from input images. The COLD approach used for representing the spatial relationship between candidate pixels in 2D and 3D images. The distribution provides dense clusters at the center for 2D text images while scattered at center for 3D images. The proposed method estimates mass for extracting such observations from each ring over the COLD distribution. The extracted mass features are fed to a Neural Network classifier for the classification of 2D and 3D images. Experimental results on our and standard datasets at image and text line levels show that the proposed method outperforms the existing methods in terms of average classification rate. The results of the non-text images show that the proposed method is content-independent. The experiments on text detection and recognition show that the proposed classification is useful for improving text detection and recognition performance. However, there is some limitation as discussed in the experimental section. Our next target is to investigate new features for improving the proposed method classification.

REFERENCES

[1] W. Wang, E. Xie, X. Li, W. Hou, T. Lu, G. Yu and S. Shao, "Shape Robust Text Detection With Progressive Scale Expansion Network", In Proc. CVPR, pp. 9328-9337, 2019

[2] P. Shivakumara, R. Raghavendra, L. Qin, K. B. Raja, T. Lu and U. Pal, "A new multi-modal approach to bib number/text detection and recognition in Marathon images", Pattern Recognition, Vol. 61, 2017, pp 479-491

[3] S. Roy, P. Shivakumara, U. Pal, T. Lu and G. H. Kumar, "Delaunay triangulation based text detection from multi-view images of natural scene", Pattern Recognition Letters, 129, 2020, pp 92-100.

[4] M. Xue, P. Shivakumara, C Zhang, T. Lu, and U. Pal, "Curved text detection in blurred/non-blurred video/scene images". Multimedia Tools and Applications, 2019, pp. 1-25.

[5] S. Tian, X. C. Yin, Y. Su and H. W. Hao, "A unified framework for tracking based text detection and recognition from web videos", IEEE Trans. PAMI, 40, pp 542-554, 2018.

[6] B. Shi, M. Yang, X. Wang, P. Luy, C. Yao and X. Bai, "ASTER: An attentional scene text recognizer with flexible rectification", IEEE Trans. PAMI, 41, pp 2035-2048, 2019.

[7] C. Luo, L. Jin and Z. Sun, "MORAN: A multi-object rectified attention network for scene text recognition", Pattern Recognition, 90, pp 109-118, 2019.

[8] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao and J. Yan, "FOTS: Fast Oriented Text Spotting with a Unified Network", In Proc. CVPR, pp 5676-5685, 2018.

[9] Y. Xu, Y. Wang, W. Zhou, Y. Wang, Z. Yang and X. Bai, "TextField: Learning a deep direction field for irregular scene text detection", IEEE Trans. IP, pp 5566-5579, 2019.

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

[11] M. Villamizar, O. Canevert and J. M. Odobez, "Multi-scale sequential network for semantic text segmentation and localization", Pattern Recognition Letters, 129, pp 63-69, 2020.

[12] S. Wang, Y. Liu, Z. He, Y. Wang and Z. Tang, "A quadrilateral scene text detector with two-stage network architecture", Pattern Recognition, 2020.

[13] M. Liao, Z. Wan, C. Yao, K. Chen and X. Bai, "Real-time scene text detection with differentiable bianrization", In Proc. Proc. AAAI, 2020.

[14] J. Xu, P. Shivakumara, T. Lu, C. L. Tan and S. Uchida, "A new method for multi-oriented graphics-scene-3D text classification in video", Pattern Recognition, pp 19-42, 2016.

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

[16] L. Nandanwar, P. Shivakumara, A. Kumar, T. Lu, U. Pal and D. Lopresti, "A new common points detection based method for classifcaiton of 2D and 3D texts in video/scene inages, In Proc. DAS, 2020.

[17] J. Xie, Z. Xiong, Q. Dai, X. Wang and Y.Zhang, "A local-gravitation-based method for the detection of outliers and boundary points", Knowledge-Based Systems, 2019.

[18] S. He and L. Schomaker, "Beyond OCR: Multi-faceted understanding of handwritten document characteristics", Pattern Recognition, 2017, pp 321-333.

[19] K. M. Ting, G. T. Zhou, F. T. Liu and S.C. Tan,"Mass estimation", pp 127-160, 2013.

[20] S. Narayan, "The Generalized Sigmoid Activation Function: Competitive Supervised Learning", Information Sciences, pp 69-82, 1997

[21] G. E. Nasr, E. A. Badr and C. Joun, "Cross Entropy Error Function in Neural Networks: Forecasting Gasoline Demand", In Proc. FLAIRS, pp 381-384, 2002.

[22] P. D. Kingma and J. L. Bai, "Adam: A method for stochastic optimization", In Proc. ICLR, pp 1-15, 2015.

[23] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C Chen , "Mobilenetv2: Inverted residuals and linear bottlenecks. In Proc. IEEE conference on computer vision and pattern recognition" (pp. 4510-4520), 2018.

[24] N. Ali and B. Zafar, "15-Scene Image Dataset". Figshare, 2018.https://doi.org/10.6084/m9.figshare.7007177.v1

[25] Ruder, Sebastian. "An overview of gradient descent optimization algorithms." arXiv preprint arXiv:1609.04747 (2016).