Recent Trends and Techniques in Text Detection and Text Localization in a Natural Scene: A Survey

Vijay Prasad[1], Pranab Das[2]

[1]Vijay Prasad
Assam Don Bosco University, Guwahati
vijay@dbuniversity.ac.in
[2]Pranab Das,
Assam Don Bosco University,Guwahati
pranab.das@dbuniversity.ac.in

Abstract*: *Text information extraction from natural scene images is a rising area of research. Since text in natural scene images generally carries valuable details, detecting and recognizing scene text has been deemed essential for a variety of advanced computer vision applications. There has been a lot of effort put into extracting text regions from scene text images in an effective and reliable manner. As most text recognition applications have high demand of robust algorithms for detecting and localizing texts from a given scene text image, so the researchers mainly focus on the two important stages text detection and text localization. This paper provides a review of various techniques of text detection and text localization.

***Keywords*: About four key words separated by commas.**

(Article history: Received: 17th April 2021 and accepted 26th June 2021)

I. INTRODUCTION

In recent years, because of widespread use of devices for taking digital images like smartphones, digital cameras, CCTVs, content-based image processing methods have gotten a lot of attention. Of all the contents in images, text information has piqued interest because it is simple to understand for both humans and computers and has a wide range of applications.

Text information is available everywhere of our live, whether that be in a roadside hoarding, or in a supermarket or in a parking lot or railway stations, airports, highways etc. It may be in a complex or simple surroundings, somewhere it may be clearly noticeable and somewhere it may be less noticeable. Presence of textual information in both artificial and natural environments, orientation or quality of text makes text extraction more complicated. Therefore, text extraction from natural scene images has led to technological developments in this area.

The textual evidence provided by text in images can be very helpful in describing the image content. Generally, text in images can be grouped into two classes, scene text and caption text. A scene text is a section of an original image when captured using mobile or digital camera, whereas a caption text is the result of artificial text being put over the original image in a later phase. In indexing and retrieving images or videos, caption text is used. The Fig.1.1 (a) shows an example of scene text and Fig.1.1 (b) shows example of caption text present in the image.

(a)

(b)

Fig. 1.1 (a) Scene Text (b) Caption Text

Text typically conveys useful information in scene images, so detecting and recognizing scene text has been considered important for a number of advanced computer vision applications such as multilingual translation, translation of street signs, image and video indexing, text recognition to help visually impaired people, industrial automation and license plate recognition etc. Since most text recognition applications require that text in images be localised in advance, robust and effective algorithms for detecting and localising texts from a given scene text image are in high demand.

The text information in the image can be extracted using the Text Information Extraction (TIE) [1] architecture as shown in Fig: 1.2. According to the architecture, there are four fundamental stages to text recognition: (i) the text detection process detects whether or not there is text in the image, (ii) Text localization refers to the process of determining the position of a text area and drawing a bounding box around it, (iii) text enhancement stage removes noise from the image and the segmentation stage segments the text i.e. divides text into meaningful units such as characters, words and sentences, and (iv) the text recognition stage uses an OCR engine to identify the text in the image.

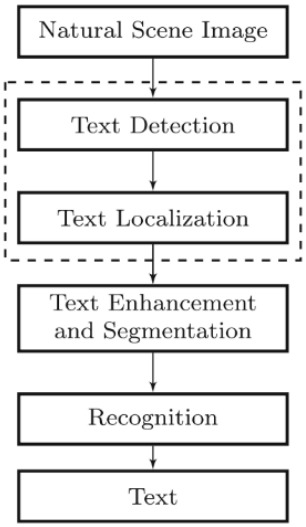

Fig 1.2: Text recognition architecture

Because of the variety of texts and the complexity of the contexts, text localization is a difficult process. Text strings, for example, may be of different sizes and fonts, moreover the orientation of the texts can be either in a straight line or can be slanted. It can also be multi-colored or curved or vertical. Text can also be affected by artefacts like clutter, distortions, low contrast or by lighting differences such as non-uniform illuminations and shadows.

The remainder of the paper is divided into the sections below. A brief overview on different benchmark dataset used in scene text detection and localization is discussed in Sections II. Section III provides a brief survey of relevant recent works on text detection and text localization. Section IV provides a brief survey of relevant works on text enhancement. Text recognition is discussed in Section V. Evaluation protocols for measuring the recognition accuracy is discussed in Section VI. And finally Section VII concludes the paper.

II. Benchmark Dataset

A number of Benchmark dataset has been used in scene text detection over the years. These datasets where used for the tasks of text detection, text localization, text segmentation and text recognition, a summary of it is shown in table 1.

III. Text Detection and Text Localization

Although text recognition has several applications, the main purpose is to identify if there is text in a given image and, if so, to detect, localize, and recognize it. Generally a scene text detection method usually consists of four successive phases these are candidate text detection, false candidate text removal, extraction and verification of text [2]–[5].

In the literature, the text recognition phases are referred to by different names as text localization [6], which aims to locate the position of candidate text, text detection which determines presence of text using the localization and verification techniques, and textual information extraction [7], [8], which is based on localization and binarization. Text enhancement on the other hand are used to improve image resolution and fix distorted text prior to the task of recognition. Also analysis of images in natural scene is mentioned in scene text recognition [9] and text recognition in the wild [10]. Therefore, for a system to recognize textual information from a natural scene text detection, localization and recognition forms the essential phases.

Techniques for text localization can be divided into two groups: region-based and texture-based techniques. Region base technique is further classified into connected component based techniques, edge based techniques and stroke based techniques. Fig 1.3 shows the categories of text localization techniques.

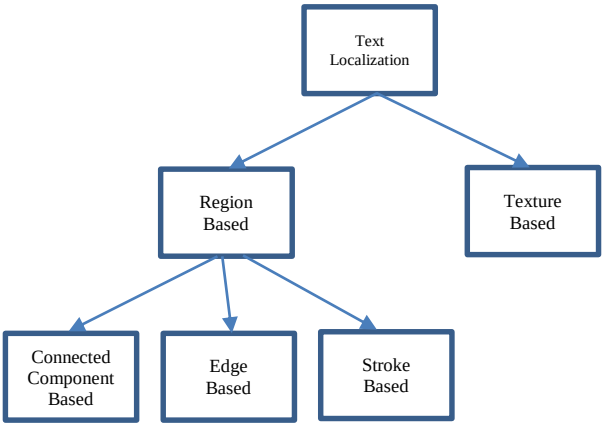

Fig 1.3: Text Localization Methods

Texture-based approaches [6][12][13][14] work under the basis that text regions have a distinct texture that allows them to be distinguished from the context or background. The texture features of the text region can be detected by discrete cosine transform (DCT) [15], histogram of oriented gradients (HOG) [16], local binary pattern [17] and wavelet transform. Texture-based techniques are relatively less sensitive to background colors; they however fail to differentiate between the text and the text like backgrounds. Effectiveness of such methods comes when the characters are dense, but the methods are slow and gives poor results if the variation in text

TABLE I. BENCHMARK DATASET

Dataset	No. of Images	Training Images, Test Images	Training words/Test Words	Labeled Words	Resolution	Type	Language	Task
ICDAR'13	462, 551	(229,233) (410,141)	848/1,095 3,564/1,439	-	-	Scene Text, Graphic text	English	Text Detection, Segmentation , Recognition
ICDAR'11	484,522	(229/255) (420/102)	(848/1,189) (3,583/918)	-	-	Scene Text, Graphic text	English	Text Detection, Segmentation , Recognition
KAIST (2010)	3000	-	-	-	640x480	Indoor, Outdoor	Korean, English, Mixed (Korean + English)	Text localization, segmentation and recognition
SVT (2010)	350	100,250	211,514	725	-	-	English	Text Localization
COCO-Text (2016)	63,686	43686, 20000	-	173589	-	Machine printed and handwritten text	English, non-English	Text localization and recognition
Synthetic Word Dataset (Oxford, VGG) (2014)	9000000	-	-	90000	-	-	-	Text recognition, segmentation
SynthText in the Wild Dataset (2016)	858750	-	-	7266866 words, 28971487 characters	-	-	-	Text detection
MSRA-TD500 (2012)	500	300,200	-	-	1296x864 1920x1280	signs, doorplates, caution plates, guide boards, billboards	English, Chinese	Text detection
IIIT 5K-Words (2012)	5000	2000,3000	-	500000	-	-	English	Text recognition
Chars74k 2009	74000	-	-	-	-	Natural images, hand drawn characters, synthesized	English, Kannada	Text recognition
MLT (2017)	18000	-	-	-	-	street signs & advertisement boards, user's photos in microblogs shops names	English, Chinese, French, Italian, Korean, German, , Japanese, Arabic and Bangla	Text detection, script identification
Street View House Numbers (SVHN)	600,000	73257, 26032 digits	-	-	32x32	House number digits	-	Number Recognition

alignment is more. The downside of such methods is their complexity, whereas they are more robust when dealing with complex background.

The Connected-component based methods [18][19][20] are based on bottom up approach that is it considers text as a set of distinct connected components having uniform colors and fulfill certain shape, size and spatial alignment constraints, distinct color and intensity. It groups the smaller components into larger ones before all of the text regions are found. In these methods connected components are extracted using Edge detection methods and color clustering methods. CC based methods are comparatively simpler to implement than edge based methods. CC based methods have lower computation cost however shows poor result in localizing text where the image background is complex. In CC based methods, some heuristic rules for text position are used to get good results. The maximal stable extremal regions (MSER) [21] is one such approach in CC based method. Advantage of MSER is that it is sensitive to image blur [21][22]. Since the number of segmented components in CC-based approaches is comparatively low, recognition can be used directly for the located text.

The edge based methods [23][24] effectively extracts text from natural scene images. These methods take into account strength of the edge, its change in orientation and edge density. These methods merge edges of the text boundary and filter out the non-text regions using several other techniques. The text regions are detected assuming that the context i.e. background is more scattered than the text. Although these methods show good results for images exhibiting large & symmetric gradient but is not very effective to detect texts with blurred image, images with large font size and images with the challenging background.

Stroke based methods[3][25][26] uses stroke feature for text detection. Each text in the image is modelled using some number of strokes, each having a different orientation. Mostly, candidate text stroke is extracted using segmentation and verification of the stroke is accomplished by feature extraction and classification. Finally, it is grouped as a single unit with the help of clustering. Due to the simplicity of the method, it is quite easy to implement. However, the method suffer in performance in complex background as segmentation and verification of text stroke is hard in such environment.

IV. Text Enhancement

Text recognition in natural scene is affected by some of the key factors like blurring, low resolution or degraded text. In such cases text enhancement techniques can be implemented to improve the recognition accuracy of the text in context.

Text enhancement uses some learning [27] or reconstruction techniques[28],[29]to improve text resolution or recover degraded text. Among conventional methods deconvolution is used for deblurring. However, it has been seen that deconvolution does not preserve the properties of the text region in the image.

In [30], an improved deconvolution technique was introduced. Here an iterative optimization algorithm based on strokes of similar widths and uniform colours has shown good results.

[27] used multi-resolution histogram encoding to model the relationship between the degraded images and their high-resolution training counterparts to generate high resolution text.

In [28], broken characters in a video is reconstructed using stroke property. Here normalized gradient features along with Canny edge map is used to extract character contours. Then a ring radius transform text enhancement technique was applied to identify the missing pixels to map the inner and outer contours of the broken character. And thus reconstructed. Text enhancement base on Sparse reconstruction has also been investigated [29][31].

V. Text Recognition

In recent years' a huge amount of textual information is being retrieved from digital images and scanned documents and thus resulting in text recognition as one of the most important area of research. Text recognition is challenging as it involves recognition of text in images which are of different fonts, size, orientation and with different background. Also presence of noise also affects the recognition performance. The text recognition process starts with capturing the image followed by preprocessing like noise removal and extraction of desired section of the text and then segmentation to extract the text region present in it. And finally text recognition converts image regions into strings. Text recognition deals with two recognition types firstly character recognition and secondly word recognition. Word recognition has been significant to text recognition as there are well-defined statistical models implementing low and high-level language features. It was comparatively easy to recognizing text at word and character levels than at the sentence level, as sentences are less tractable than words.

A. Character Recognition

In character recognition where characters are in the same font, linear discriminant analysis and Gabor feature and is used [26]. In case where characters are in different fonts or are distorted, classifying is difficult because of diversity within a class [32]. However specific classifier for each of them can be used to sort such problems [32][33].

Other ways of character recognition are done using unsupervised learning [34], image rectification algorithms[35], or feature pooling [36].

In literature some of the methods of character recognition include template matching[37] or matrix matching. Here all the character images are kept as templates. Character

classification is achieved by matching score of the input character image against the templates. The score results in a similarity measure. Higher value of similarity measure means that the probability of match between character image and template is on a higher side. Also certain variation of the existing techniques is used for classification. One such way is to use wavelet transform of the character image and template for matching instead of pixel based approach.

Structural classification [37] techniques uses structural features such as character strokes, character holes, character corners and character concavities to classify characters. Here structural features of the input text character are extracted and sent to a rule-based system to identify the class of the character in context. Instead of using rule based system nearest neighbor classifier is also used as a variation in the feature space.

In some other methods discriminant function classifiers operate in multi-dimensional feature spaces to distinguish character feature descriptions and minimise mean-squared classification error. Bayesian classifiers uses probability theory to minimize loss function related with character misclassification. Artificial Neural Networks based character classification techniques that uses mathematical optimization to minimize classification errors.

B. Word Recognition

Word recognition in natural scene images is a different form of Optical character recognition. ICDAR dataset available for natural scenes has been used in the recent years for Robust Reading Competitions. This has led to extensive work on word/text recognition. Scene Text Recognition is still a challenging task because of multiple reasons, such as complex backgrounds, multi oriented text, different fonts, and inadequate imaging environment. Fig 1.4 shows Complex, Multiple Font, Multi oriented text.

Fig: 1.4: Complex, Multiple Font, Multi oriented text

In word recognition it is observed that the recognition model fails to recognize word/text which are either distorted or degraded. The recognition model may classify identical characters as different due such distortions and unavailability of sufficient training data for specific fonts [9].

Also language models can be integrated with character recognition using some optimization techniques such as Bayesian inference [9][38], graph models[34], Markov [39], Conditional random fields [40].

In [9] a model is constructed by combining key information such as similarity to other characters, appearance, a lexicon and language. Similarity function was integrated for improved recognition in case of lesser font samples. It incorporated a lexicon into the model and removed recognition errors and improved accuracy.

In [41] a bottom-up (character) and top-down (language) framework was used for text recognition. A sliding window classification was used for character detections, along with that a CRF model was used to strengthen the detections. Also higher order language models (n-grams) and maximum a posteriori (MAP) have been used to improve recognition accuracy. Large dictionaries to assist weak character recognition and non-dictionary words has been adopted to implement a high order language model

VI. Evaluation Protocol

The performance of any text recognition technique is measured using precision, recall and F1-score. These matrices are computed based on the parameters of confusion matrix. The parameters of confusion matrix are computed based on the predicted text instances which is matched against the ground truth.

True Positives (TP) - It means value of both actual and predicted class is positive.

True Negatives (TN) - It means value of both actual and predicted class is negative.

False Positives (FP) - It means value of actual class is negative and value of predicted class is positive.

False Negatives (FN) - It means value of actual class is positive and value of predicted class is negative.

A. Precision

The term "precision" refers to how accurate or precise a model is. How many of the cases that were predicted to be positive really turned out to be positive? It is also called a measure of positive predictive value.

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)}$$

B. Recall

The number of real positives that the model predicted as positives is calculated by recall. It is also known as True positive rate (TPR) or sensitivity.

$$Recall = \frac{True\ Positive}{(True\ Positive + False\ Negative)}$$

C. F1 score

F1 Score is the harmonic mean of Recall and Precision

$$F1\ Score = \frac{2\ X\ (Recall\ X\ Precision)}{(Recall + Precision)}$$

Here a summary of the detection accuracy in terms of Precision, Recall and F1 score of various state-of-the-art approaches on the widely used scene text datasets ICDAR *2013 and* ICDAR *2015* are shown in *Table 2* and *Table 3.*

TABLE 2: Scene Text Detection on ICDAR-2013

Method	**Precision**	**Recall**	**F1**
Zhang et al. [42]	88	78	83
CTPN [43]	93	83	88
Holistic [44]	88.88	80.22	84.33
PixelLink [45]	86.4	83.6	84.5
He et al. [46]	92	80	86
EAST [47]	92.64	82.67	87.37
SSTD [48]	89	86	88
Lyu et al. [49]	93.3	79.4	85.8
Liu et al. [50]	88.2	87.2	87.7
WordSup [51]	93.34	87.53	90.34
Lyu et al.[52]	94.1	88.1	91.0
Baek et al. [53]	97.4	93.1	95.2

TABLE 3: Scene Text Detection on ICDAR-2015

Method	**Precision**	**Recall**	**F1**
Zhang et al. [42]	71	43.0	54
CTPN [43]	74	52	61
Holistic [44]	72.26	58.69	64.77
PixelLink [45]	85.5	82.0	83.7
He et al. [46]	82	80	81
EAST [47]	83.57	73.47	78.20
SSTD [48]	80	73	77
Lyu et al. [49]	94.1	70.7	80.7
Liu et al. [50]	72	80	76
WordSup [51]	79.33	77.03	78.16
Lyu et al.[52]	85.8	81.2	83.4
Baek et al. [53]	89.8	84.3	86.9

VII. Conclusion

A study of current text detection, localization, enhancement, and text recognition techniques in natural scene images was presented in this paper. Despite the fact that much work has been done on text recognition and localization in the past, there is still a lot of scope to create a versatile system that can work in a variety of environments and text formats. In most text detection systems designed for natural scene, extraction of the text regions has always remained a challenging task. With the rapid growth of machine learning and other state of the art techniques, design of new, effective and robust techniques which can detect and localize text will be an important task.

References

[1] K. Jung, K. I. Kim, and A. K. Jain, 'Text information extraction in images and video: A survey', *Pattern Recognit.*, vol. 37, no. 5, pp. 977–997, 2004, doi: 10.1016/j.patcog.2003.10.012.

[2] W. Huang, Z. Lin, J. Yang, and J. Wang, 'Text localization in natural images using stroke feature transform and text covariance descriptors', in *Proceedings of the IEEE International Conference on Computer Vision*, IEEE, Dec. 2013, pp. 1241–1248. doi: 10.1109/ICCV.2013.157.

[3] B. Epshtein, E. Ofek, and Y. Wexler, 'Detecting text in natural scenes with stroke width transform', in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2010, pp. 2963–2970. doi: 10.1109/CVPR.2010.5540041.

[4] C. Yao, X. Bai, and W. Liu, 'A unified framework for multioriented text detection and recognition', *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4737–4749, Nov. 2014, doi: 10.1109/TIP.2014.2353813.

[5] X. C. Yin, X. C. Yin, K. Huang, and H. W. Hao, 'Robust text detection in natural scene images', *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 970–983, May 2014, doi: 10.1109/TPAMI.2013.182.

[6] R. Lienhart and A. Wernicke, 'Localizing and segmenting text in images and videos', *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 4, pp. 256–268, Apr. 2002, doi: 10.1109/76.999203.

[7] S. Uchida, 'Text localization and recognition in images and video', in *Handbook of Document Image Processing and Recognition*, London: Springer London, 2014, pp. 843–883.

[8] J. Zhang and R. Kasturi, 'Extraction of Text Objects in Video Documents: Recent Progress', in *2008 The Eighth IAPR International Workshop on Document Analysis Systems*, 2008, pp. 5–17. doi: 10.1109/DAS.2008.49.

[9] J. J. Weinman, E. Learned-Miller, and A. R. Hanson, 'Scene text recognition using similarity and a lexicon with sparse belief propagation', *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 10, pp. 1733–1746, Oct. 2009, doi: 10.1109/TPAMI.2009.38.

[10] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng, 'End-to-end text recognition with convolutional neural networks', in *Proceedings of the 21st International*

Conference on Pattern Recognition (ICPR2012), Nov. 2012, pp. 3304–3308.

[11] X. Chen and A. L. Yuille, 'Detecting and reading text in natural scenes', in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, 2004. doi: 10.1109/cvpr.2004.1315187.

[12] J. Gllavata, R. Ewerth, and B. Freisleben, 'Text detection in images based on unsupervised classification of high-frequency wavelet coefficients', in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, Aug. 2004, pp. 425-428 Vol.1. doi: 10.1109/ICPR.2004.1334146.

[13] Huiping Li, D. Doermann, and O. Kia, 'Automatic text detection and tracking in digital video', *IEEE Trans. Image Process.*, vol. 9, no. 1, pp. 147–156, Jan. 2000, doi: 10.1109/83.817607.

[14] Q. Ye, W. Gao, and D. Zhao, 'Fast and robust text detection in images and video frames', *Image Vis. Comput.*, vol. 23, pp. 565–576, 2005, doi: 10.1016/j.imavis.2005.01.004.

[15] H. Zhang, K. Zhao, Y.-Z. Song, and J. Guo, 'Text extraction from natural scene image: A survey', *Neurocomputing*, vol. 122, pp. 310–323, Dec. 2013.

[16] J. Zhang and R. Kasturi, 'Text Detection Using Edge Gradient and Graph Spectrum', in *2010 20th International Conference on Pattern Recognition*, Aug. 2010, pp. 3979–3982. doi: 10.1109/ICPR.2010.968.

[17] B. Bai, F. Yin, and C. L. Liu, 'Scene text localization using gradient local correlation', in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2013. doi: 10.1109/ICDAR.2013.279.

[18] Y. Zhong, K. Karu, and A. K. Jain, 'Locating text in complex color images', *Pattern Recognit.*, vol. 28, no. 10, pp. 1523–1535, 1995, doi: https://doi.org/10.1016/0031-3203(95)00030-4.

[19] V. Y. Mariano and R. Kasturi, 'Locating uniform-colored text in video frames', in *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, 2000, pp. 539–542 vol.4. doi: 10.1109/ICPR.2000.902976.

[20] H. Wu, B. Zou, Y. Zhao, and J. Guo, 'Scene text detection using adaptive color reduction, adjacent character model and hybrid verification strategy', *Vis. Comput.*, vol. 33, no. 1, pp. 113–126, 2017, doi: 10.1007/s00371-015-1156-1.

[21] J. Matas, O. Chum, M. Urban, and T. Pajdla, 'Robust wide-baseline stereo from maximally stable extremal regions', *Image Vis. Comput.*, vol. 22, no. 10 SPEC. ISS., pp. 761–767, Sep. 2004, doi: 10.1016/j.imavis.2004.02.006.

[22] Y. Li and H. Lu, 'Scene text detection via stroke width', 2012, pp. 681–684.

[23] C. Yu, Y. Song, and Y. Zhang, 'Scene Text Localization Using Edge Analysis and Feature Pool', *Neurocomput.*, vol. 175, no. PA, pp. 652–661, Jan. 2016, doi: 10.1016/j.neucom.2015.10.105.

[24] S. Lee, M. S. Cho, K. Jung, and J. H. Kim, 'Scene Text Extraction with Edge Constraint and Text Collinearity', in *2010 20th International Conference on Pattern Recognition*, Aug. 2010, pp. 3983–3986. doi: 10.1109/ICPR.2010.969.

[25] C. Yi and Y. Tian, 'Text Detection in Natural Scene Images by Stroke Gabor Words', in *2011 International Conference on Document Analysis and Recognition*, 2011, pp. 177–181. doi: 10.1109/ICDAR.2011.44.

[26] Y. Pan, Y. Zhu, J. Sun, and S. Naoi, 'Improving Scene Text Detection by Scale-Adaptive Segmentation and Weighted CRF Verification', in *2011 International Conference on Document Analysis and Recognition*, 2011, pp. 759–763. doi: 10.1109/ICDAR.2011.158.

[27] G. Caner and I. Haritaoglu, 'Shape-DNA: Effective Character Restoration and Enhancement for Arabic Text Documents', in *Proceedings of the 2010 20th International Conference on Pattern Recognition*, in ICPR '10. USA: IEEE Computer Society, 2010, pp. 2053–2056. doi: 10.1109/ICPR.2010.506.

[28] P. Shivakumara, T. Q. Phan, S. Bhowmick, C. L. Tan, and U. Pal, 'A Novel Ring Radius Transform for Video Character Reconstruction', *Pattern Recogn.*, vol. 46, no. 1, pp. 131–140, Jan. 2013, doi: 10.1016/j.patcog.2012.07.008.

[29] Y. Lou, A. L. Bertozzi, and S. Soatto, 'Direct Sparse Deblurring', *J. Math. Imaging Vis.*, vol. 39, no. 1, pp. 1–12, 2011, doi: 10.1007/s10851-010-0220-8.

[30] H. Cho, J. Wang, and S. Lee, 'Text Image Deblurring Using Text-Specific Properties', in *Computer Vision -- ECCV 2012*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 524–537.

[31] C. Yao, X. Zhang, X. Bai, W. Liu, Y. Ma, and Z. Tu, 'Rotation-Invariant Features for Multi-Oriented Text Detection in Natural Images', *PLoS One*, vol. 8, no. 8, p. e70173, Aug. 2013, doi: 10.1371/journal.pone.0070173.

[32] K. Sheshadri and S. Divvala, 'Exemplar Driven Character Recognition in the Wild', 2012, doi: 10.5244/C.26.13.

[33] K. Kuramoto, W. Ohyama, T. Wakabayashi, and F. Kimura, 'Accuracy Improvement of Viewpoint-Free Scene Character Recognition by Rotation Angle Estimation', in *Revised Selected Papers of the International Workshop on Camera-Based Document Analysis and Recognition - Volume 8357*, Berlin, Heidelberg: Springer-Verlag, 2013, pp. 60–70. doi: 10.1007/978-3-319-05167-3_5.

[34] A. Coates *et al.*, 'Text detection and character recognition in scene images with unsupervised feature learning', in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2011. doi: 10.1109/ICDAR.2011.95.

[35] J. Liu, H. Li, S. Zhang, and W. Liang, 'A novel italic detection and rectification method for Chinese advertising images', in *Proceedings of the*

International Conference on Document Analysis and Recognition, ICDAR, IEEE, Sep. 2011, pp. 698–702. doi: 10.1109/ICDAR.2011.146.

[36] C. Y. Lee, A. Bhardwaj, W. Di, V. Jagadeesh, and R. Piramuthu, 'Region-based discriminative feature pooling for scene text recognition', in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2014, pp. 4050–4057. doi: 10.1109/CVPR.2014.516.

[37] J. Parker, *Algorithms for Image Processing and Computer Vision*. 1997.

[38] D. Zhang and F. Chang, 'A Bayesian framework for fusing multiple word knowledge models in videotext recognition', 2003, pp. II–528. doi: 10.1109/CVPR.2003.1211512.

[39] J. Weinman, E. Learned-Miller, and A. Hanson, 'A Discriminative Semi-Markov Model for Robust Scene Text Recognition', in *IEEE, Proc. Intl. Conf. on Pattern Recognition (ICPR*, 2008, pp. 1–5. doi: 10.1109/ICPR.2008.4761818.

[40] C. Shi, C. Wang, B. Xiao, Y. Zhang, S. Gao, and Z. Zhang, 'Scene text recognition using part-based tree-structured character detection', in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2013, pp. 2961–2968. doi: 10.1109/CVPR.2013.381.

[41] A. Mishra, K. Alahari, and C. V. Jawahar, 'Top-down and bottom-up cues for scene text recognition', *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 2687–2694, 2012, doi: 10.1109/CVPR.2012.6247990.

[42] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai, 'Multi-oriented Text Detection with Fully Convolutional Networks', in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2016, pp. 4159–4167. doi: 10.1109/CVPR.2016.451.

[43] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, 'Detecting Text in Natural Image with Connectionist Text Proposal Network', *CoRR*, vol. abs/1609.03605, 2016.

[44] C. Yao, X. Bai, N. Sang, X. Zhou, S. Zhou, and Z. Cao, 'Scene Text Detection via Holistic, Multi-Channel Prediction', *CoRR*, vol. abs/1606.09002, 2016.

[45] D. Deng, H. Liu, X. Li, and D. Cai, 'PixelLink: Detecting scene text via instance segmentation', *32nd AAAI Conf. Artif. Intell. AAAI 2018*, vol. 32, no. 1, pp. 6773–6780, Apr. 2018, doi: 10.1609/aaai.v32i1.12269.

[46] W. He, X. Y. Zhang, F. Yin, and C. L. Liu, 'Deep Direct Regression for Multi-oriented Scene Text Detection', *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2017-Octob, pp. 745–753, Mar. 2017, doi: 10.1109/ICCV.2017.87.

[47] X. Zhou *et al.*, 'EAST: An efficient and accurate scene text detector', in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, pp. 2642–2651. doi: 10.1109/CVPR.2017.283.

[48] P. He, W. Huang, T. He, Q. Zhu, Y. Qiao, and X. Li, 'Single Shot Text Detector with Regional Attention', *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2017-Octob, pp. 3066–3074, Aug. 2017, doi: 10.1109/ICCV.2017.331.

[49] P. Lyu, C. Yao, W. Wu, S. Yan, and X. Bai, 'Multi-oriented Scene Text Detection via Corner Localization and Region Segmentation', in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2018, pp. 7553–7563. doi: 10.1109/CVPR.2018.00788.

[50] Z. Liu, G. Lin, S. Yang, J. Feng, W. Lin, and W. L. Goh, 'Learning Markov Clustering Networks for Scene Text Detection', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018.

[51] H. Hu, C. Zhang, Y. Luo, Y. Wang, J. Han, and E. Ding, 'WordSup: Exploiting Word Annotations for Character Based Text Detection', *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2017-Octob, pp. 4950–4959, Aug. 2017, doi: 10.1109/ICCV.2017.529.

[52] M. Liao, P. Lyu, M. He, C. Yao, W. Wu, and X. Bai, 'Mask TextSpotter: An End-to-End Trainable Neural Network for Spotting Text with Arbitrary Shapes', *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 532–548, Feb. 2021, doi: 10.1109/TPAMI.2019.2937086.

[53] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, 'Character Region Awareness for Text Detection', *CoRR*, vol. abs/1904.01941, 2019.

AUTHOR PROFILE

Mr. Vijay Prasad
Assam Don Bosco University, Assistant Professor (II) and with 10 years of teaching and research experience. Area of specialization: Computer vision and pattern recognition and Web technologies. Detailed profile can be accessed at https://sites.google.com/view/vpdvijay

Dr. Pranab Das
Assam Don Bosco University, Assistant Professor (II) and with 12 years of teaching and research experience. Area of specialization: Speech Processing, Image Processing, Machine Learning, Audio Mining. Detailed profile can be accessed at http://erp.dbuniversity.ac.in/emplist/viewprofile.php?id=87