



Character Recognition in Scene Images Using MSER and CNN

R. P. Rajeswari¹(✉) and B. Aradhana²

¹ Rao Bahadur Y Mahabaleswarappa Engineering College, Ballari, Karnataka 583104, India
rajeswarirp@rymec.in

² Ballari Institute of Technology and Management, Ballari, Karnataka 583104, India

Abstract. Text detection in Scene images has procured significance in recent decade. Due to its diversified applications in blind navigation assistance for Visually impaired, traffic monitoring, Automatic driving assistance systems etc., Text detection has stimulated new research avenues in area of computer vision Text detection is a trivial task because of varying color, font face and size, orientation of text against complex background. A diversity of deep learning techniques are introduced by researchers for graphical text detection in images. The article proposed method consisting of 3 stages. First, we use Otsu's method for text separation from background. Secondly Text ROI's are extracted using Maximally stable Ensemble method (MSER). Finally, each extracted text ROI is classified using ConvNets. CNN classifier have been trained to recognize Scene Text Characters.

Keywords: Text detection · Classification · MSER · CNN

1 Introduction

Advances in Technology have lead to increased usage of smart phones, tablets and digital cameras, resulting in large collection of heterogeneous data consisting of video images, natural scene images and web based images with text. These images contain useful text that can be used for numerous applications such as machine language translation, safe vehicle driving, tracking and recognition of license plate, spot identification, house number tracking from maps, image retrieval, intelligent transportation etc. Text Reading in wild is major research issue. Scene text reading promotes significant clues for content-based retrieval applications. Objective of detection is to locate region of text in image. The text recognition identifies and generates text from these images. In other words, Text detection task is to find a minimum sized region of interest with all of text in the image inside it. Text Detection and Recognition determines text areas using bounding boxes in an image and output a sequence of characters associated with its content. Characteristics of scene Text: Style/Size - Text in images appears either in printed block letters or in handwritten/Calibrated cursive form with varying size. Spacing – Spacing between characters and words together with the size of text causes detection difficult. Color-Text in scene images can have multiple font color. Background- Scene images have complex background with embedded text and sometimes get merged with background. Hence text detection against complex background with low resolution is challenging (Fig. 1).



Fig. 1. Scene Text images with variation in background, multi font color, orientation [24, 25]

In recent past, researchers have proposed several text detection approaches for images and video frames. Due to variation in text color, font face, multiple orientations of text, complex background, and geometric distortions in images, there is tendency of failing to detect true text regions. Further, the growth in Optical Character Recognition (OCR) systems has made computers to read text from images. Since Images may have many other non-character textures, it is difficult for the OCR to read text. We need to extract character strings from images.

Text detection go through three phases: The first phase is to detect presence of text in scene images. Second step consists of localizing text or finding the regions of text in scene image, the third phase is text recognition, which transforms the detected text into transcription. The primary objective of text detection in wild is to generate bounding boxes for diverse text blocks in image.

2 Literature Survey

Text embedded in images is rich source of semantic information which is extracted and used for a variety of applications. Scene Text detection has evolved as active area of research in area of Computer Vision & Deep Learning. Numerous state-of- art approaches and models are introduced by the researchers of computer vision and scene text community. Based on the existing literature review, Detection of graphical text can be largely categorized into traditional text detectors that uses hand crafted features to detect text in images and secondly deep learning based methods. Traditional methods are classified into sliding window (SW) and connected component (CC) based methods.

Sliding Window (SW): In SW approach, a small window is slide over entire scene image. A classifier with predefined feature set used to find the occurrence of text in images. Raghunath Roy et al. [1] used the SW method with edit distance for handwritten text recognition on MNIST dataset. Wang et al. [2] used sliding window method with CNN to identify candidate text lines in scene images. Mishra et al. [3], applied SW method with aspect ratio of each character to obtain locations of text in scene image.

Connected Component (CC): These extract candidate components of text from image. Further trained classifier with features are used to eliminate non text components. Stroke width Transform (SWT) and maximally stable extremal regions widely used CC methods. N.Gupta et al. [4] used grab cut to segment the text region followed by MSER feature detector. Rituraj Soni et al. [5] extracted text components by smoothing edges using guided filter and MSER. Juli P et al. [6], used stroke width transform (SWT) to identify text in natural scenes. Here the deskewing algorithm is sed for deskewing in order to detect text for image irrespective of its orientation, able to detect text of any

font, orientation, direction and scale Arpit Jain et al. [7], proposed an end-to-end system for text identification from videos using Maximally Stable External Regions (MSER) to detect text in very low illuminated regions and Super vector machine (SVM) classifier is used to classify the text /non text regions. Shahzia Siddiqua et al. [8], used morphological operations to identify characters of Kannada graphical text from scene images. The method consists of Edge detection, filtering of features and Binarization. It works fine for text regardless of image disparity, complex background, font size & type of text, but fails to detect smaller & dense font size.

S.A. Angadi et al. [9], proposed method that uses Profile features Zone wise to extract regions of text from mobile captured images of lower resolution and is insensitive to the font type & size variation, thickness and inter character spacing.

Hybrid Methods: To resolve the disadvantage of SW and CC methods, combination of different schemes called hybrid methods are used. Youbao Tang et al. [10], used pixel stroke feature transform and region classification to detect scene text in various intricate scenarios. Mitra Behzadi et al. [11], proposed text detection using Fully Convolution Dense Net..., which segments each image into 3 sections namely foreground text, background and word-fence. This approach works fine for limited datasets. Yirui Wu et al. [12], proposes Multi-domain Stroke Symmetry Histogram (MSSH) with deep convolution network to find text in scene images. Symmetry property represented by stroke pairs is used as in MSSH, to capture the characteristics of text Yuliang Liu, Lianwen Jim et al. [13], proposed a new CNN based Deep matching prior network (DMP Net) to spot text using quadrilateral sliding window based on multi orientation and shape of text. Quadrilateral of varying sizes are used to recall text. Overlapping threshold intrinsic used to find out whether window with sliding polygon is negative or positive, where Positive window is used to localize the text. Hence, Monte Carlo method is used for accurate computing of polygonal areas. The proposed method works efficiently for the reduction of background interference. Dao Wu et al. [14], describes framework to find text in scene images based on a Strip-based Text Detection Network (STDN), a region proposal network. Cascaded learning is used. Re Xionlang et al. [15], used Convolutional neural networks to detect text lines which is capable of learning hierarchical and discriminative features for classification WafaKhlif et al. [16], introduced detection of text in natural images consisting of CC analysis at multi-levels and CNN is used to learn the components of text. Further, graph-based grouping is done to prevent overlapping of text boxes. Lan Wang et al. [17], presented wild text detection, which significantly exploits cues of the text background regions. Specifically, text candidates and probable text background regions are extracted from the video frame. Shape, Spatial and motional correlations among graphical foreground text against complex background region are exploited with a bipartite graph model. To obtain better accuracy and to refine text candidates Random walk algorithm is used.

Deep Learning models: Deep learning methods are characterized by automated feature learning with fast and accurate text detection. Further, in recent past, State of art Deep Learning models using transfer learning have been used in text detection. These models include VGG, Resnet and Inception models. Baoguang Shi et al. [18], introduced Segment Linking as text detection method. A segment is box with orientation covering

a part of text line or word; Two adjoining segments, are connected via Links, Adjoining segments imply that they belong to same word or text line. Segmented Links are identified by fully-Convolution neural network. Seglink method proved to be efficient for horizontal, arbitrary oriented and multi-lingual text. The links connect adjoining segments, but unsuccessful to link segments with larger inter space distance.

Minghui Liao et al. [19], presented Textboxes+ +, an rapid text detector which detects text with multiple orientation in a single forward pass. Arbitrary-oriented texts are represented by tighter quadrilaterals or rectangles with arbitrary orientation. Text Boxes+ + fails to detect characters of large spacing since it is difficult to fix the precise boundary for vertical and arched texts owing to inadequate representation of quadrilaterals.

Wenhao He et al. [20], used Fast Convolutional Network (FCN) for multiple oriented and multilingual graphical text detection. One of the observed limitations is lack of detection of end words in text line. Xinyu Zhou et al. [21], proposed EAST (efficient & accurate scene text detector) which uses pipeline of fully Convolutional layer with NMS merging. Yoshito Nagaoka et al. [22], proposed text detection using Faster RCNN. Multiple Text Region proposals are obtained for each distinct convolution layer and helps in text detection with varying size [22].

3 Methodology

The proposed method is defined with following steps.

1. Background filtering using Otsu method.
2. To extract ROI of text region using MSER method.
3. To recognize extracted characters using CNN.

Figure 2 represents the pipeline of proposed text detection method.



Fig. 2. Pipeline of proposed method for character recognition in scene images

3.1 Background Filtering Using Otsu's Method

Since Scene image Text is embedded in complex background, in order to enhance the accuracy of detection, Text present in foreground with varying font size, style, orientation needs to be separated from the complex background from image. Inspired by the significant performance of Otsu's method, it is used for separating foreground text from background. Otsu method separates foreground objects and background objects by means of maximum variance. Figure 3 displays the output.

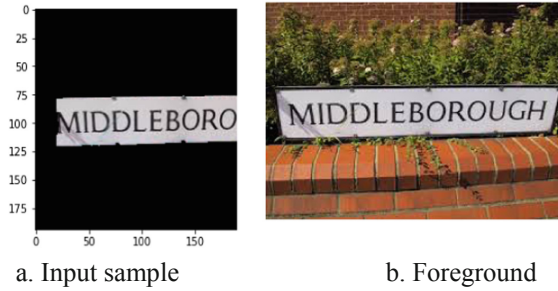


Fig. 3. Text in Foreground separated from background

3.2 Extraction of ROI from Scene Images Using MSER

Mata's et al., proposed Maximally Stable Extremal Regions (MSER) for detecting region of interest for wide-baseline stereo matching. MSER method tries to distinguish different regions in image based on properties such as affine transformation, invariance of the intensity function, a measure of stability of images etc. MSER extracts connected components and groups all pixel with same intensities, thus forming Regions. Extremal signifies the intensity levels within the MSER region differ highly with respect to the outer regions.

Algorithm: MSER to extract Text Constituents in Scene Images

Input: Sample Image

Output: Text Constituents of Sample Image.

1. Convert BGR image to Gray Scale
2. Detect Text Regions in Gray Scale Image using MSER
3. Obtain Contour for each Character of Text in the Region
4. Set each Contour in image as Region of Interest (ROI) for text extraction
Extract and save ROI's

MSER gives the ROI's of text regions as shown in Fig. 4.



Fig. 4. Sample ROI'S extracted using MSER

3.3 Character Classification Using CNN

CNN is a multilayered architecture with feature convolutions. CNN is designed to learn high level features for visual recognition. CNN consists of input layer to read graphical input image, numerous hidden layers and output layer. Input layer reads the input image

in the form of 2D array Feature Extraction is done by convolutional layers. Each convolutional layer consists of numerous kernel of size 3×3 or 5×5 , using which the image is convolved. Convolution operation is carried out on image, where Kernel slides over the input image and performs sum of product at every location is computed. Stride represents the size of each sliding step of Kernel. Slide size of 1 or 2 is considered. Feature maps produced by convolutional layers are fed as input to subsequent layers. RELU, a Non Linear activation function is applied to output of these convolutional layers, which replaces negative values with zero, thus speeding up the learning process. Pooling layer reduces the computation overhead & spatial size of feature maps. Generally, three types of Pooling is used: Min, Average and Max Pooling. Succeeding is Fully Connected layer. It executes the task of classification and prediction by learning through feature maps. Finally, SoftMax is used to output the class probabilities [23].

In the proposed method, CNN character Classifier is a sequential model. An input image of size 28×28 is given as graphical input image to input layer. Next to input layer is first 2D convolution layer with 16 filters with 3×3 Kernel size and activation function 'RELU' followed by 2×2 Max Pooling layer. Second, third and fourth layer is 2D Convolutional layer with filters 32, 64, 64 with kernel size of 3×3 respectively and subsequently followed by a Max Pooling layer of size 2×2 . Two dense layers with activations function 'RELU' and 'SOFTMAX' Model are used to compile model. The output is a character class that the character belongs to. The proposed CNN Architecture is shown in Fig. 5.

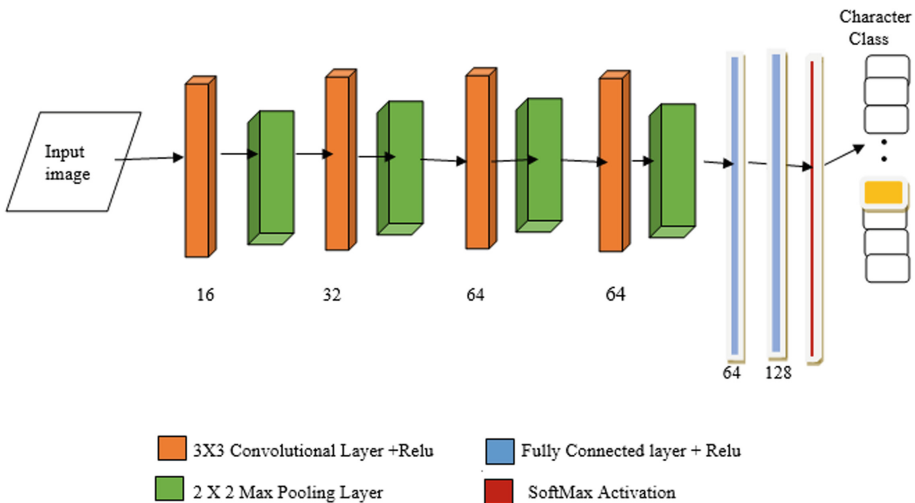










Fig. 5. Architecture of convolutional neural network for character recognition

4 Observations and Experimental Results

Scene images are drawn randomly from SVT dataset and IITK dataset Since the scene text image contains a complex background, which interferes with the process of text

detection, graphical text in foreground is separated from complex background using Otsu’s method. Secondly, text areas are detected and extracted using Maximally stable Ensemble method. Each extracted region of interest (ROI) contains character of text in image. MSER significantly detects text in images with varying colors and contrast. Further, CNN character classifier is trained using A-Z Kaggle character dataset with a train test split ratio of 80% and 20% respectively. The CNN classifier performs significantly. It is able to predict the character class of each extracted text ROI with an accuracy of 92%-97%. However, there are certain miss- predictions of characters extracted as displayed in Table 1. These miss-predictions can be eliminated further by improving the proposed method and using pretrained models.

Table 1. Samples of character classification Using CNN

Actual class: L	Actual class: B	Actual class: N	Actual class: M
Predicted class: L	Predicted class: B	Predicted class: N	Predicted class: M
			
Actual class:S	Actual class:G	Actual class:U	Actual class:R
Predicted class: S	Predicted class: J	Predicted class: M	Predicted class: K
			
Misclassification			

5 Conclusion

Localization and identification of graphical Text in natural images is solitary major research challenges in Computer Vision due to variations in text color, size, font face, type, Background with Low illuminated background with arbitrary orientation. In this paper, First foreground text is separated from background using Otsu method. Secondly, Text regions are determined based on color features using MSER algorithm. Text ROI’s are saved individually as characters images. CNN classifier is trained to predict the output class of character in image with low error rate. However there are certain misclassifications which can be resolved by using State-of-art pre-trained models.

References

1. Dey, R., Balabantaray, R.C., Mohanty, S.: Sliding window based off-line handwritten text recognition using edit distance. *Multimed. Tools Appl.* 1573–7721 (2021)

2. Wang, K., et al.: End-to-end scene text recognition. In: 2011 International Conference on Computer Vision, pp. 1457–1464 (2011)
3. Mishra, A., Alahari, K., Jawahar, C.V.: Scene text recognition using higher order language priors. In: Proceedings British Machine Vision Conference, pp. 1–11 (2012)
4. Gupta, N., Jalal, A.S.: A robust model for salient text detection in natural scene images using MSER feature detector and Grabcut. *Multimed. Tools Appl.* **78**(8), 10821–10835 (2018). <https://doi.org/10.1007/s11042-018-6613-1>
5. Soni, R., Bijendra K., Satish, C.: Text detection and localization in natural scene images using MSER and fast guided filter. In: 2017 Fourth International Conference on Image Information Processing (ICIIP). IEEE (2017)
6. Bhirud, J.P., Rege, P.P.: A modified SWT based text-image separation in natural scene images. In: Advances in Signal Processing (CASP), Conference on. IEEE (2016)
7. Jain, A., Peng, X., Zhuang, X., Natarajan, P., Cao, H.: Text detection and recognition in natural scenes and consumer videos. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp. 1245–1249. IEEE (2014)
8. Siddiqua, S., Naveena, C., Manvi, S.K.: A combined edge and connected component based approach for Kannada text detection in images. In: 2017 International Conference on Recent Advances in Electronics and Communication Technology (ICRAECT). IEEE (2017)
9. Angadi, S.A., Kodabagi, M.M.: A light weight text extraction technique for hand-held device. *Int. J. Image Graph.* **15**(04), 1550017 (2015)
10. Youbao, T., Wu, X., Member, IEEE. Scene text detection using superpixel-based stroke feature transform and deep learning based region classification. *IEEE Trans. Multimed.* **20**(9), 2276–2288 (2018)
11. Behzadi, M., Safabakhsh, R.: Text detection in natural scenes using fully convolutional densenets. In: 2018 4th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS). IEEE (2018)
12. Yirui, W., Wang, W., Palaiahnakote, S., Lu, T.: A robust symmetry-based method for scene/video text detection through neural network. document analysis and recognition (ICDAR). In: 2017 14th IAPR International Conference on, vol. 1. IEEE (2017)
13. Liu, Y., Jin, L.: Deep matching prior network: toward tighter multi-oriented text detection. *Proc. CVPR.* (2017)
14. Wu, D., Wang, R., Dai, P., Zhang, Y., Cao, X.: Deep strip-based network with cascade learning for scene text localization. In: Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on, vol. 1. IEEE (2017)
15. Ren, X.: A novel scene text detection algorithm based on convolution neural network. In: Visual Communications and Image Processing (VCIP), 2016. IEEE (2016)
16. Khlif, W., Nayef, N., Burie, J.-C., Ogier, J.-M., Alimi, A.: Learning text component features via convolutional neural networks for scene text detection. In: 2018 13th IAPR International Workshop on Document Analysis Systems (DAS), pp. 79–84. IEEE (2018)
17. Wang, L., Wang, Y., Shan, S., Su, F.: Scene text detection and tracking in video with background cues. In: Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, pp. 160–168. ACM (2018)
18. Shi, B., Bai, X., Belongie, S.: Detecting oriented text in natural images by linking segments. arXiv preprint [arXiv:1703.06520](https://arxiv.org/abs/1703.06520) (2017)
19. Liao, M., Shi, B., Bai, X.: Textboxes++: a single-shot oriented scene text detector. *IEEE Trans. Image Process.* **27**(8), 3676–3690 (2018)
20. He, W., Zhang, X.-Y., Yin, F., Liu, C.-L.: Multi-oriented and multi-lingual scene text detection with direct regression. *IEEE Trans. Image Process.* **27**(11), 5406–5419 (2018)
21. Zhou, X., et al.: East: an efficient and accurate scene text detector. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2017)

22. Nagaoka, Y., Miyazaki, T., Sugaya, Y., Omachi, S.: Text detection by faster R-CNN with multiple region proposal networks. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), pp. 15–20 (2017). <https://doi.org/10.1109/ICDAR.2017.343>
23. Balakrishna, K., Rao, M.: Tomato plant leaves disease classification using KNN and PNN. *Int. J. Comput. Vision Image Process.* **9**(1), 51–63 (2019). <https://doi.org/10.4018/IJCVIP.2019010104>
24. https://shodhganga.inflibnet.ac.in/bitstream/10603/125752/6/06_chapter1.pdf
25. <https://cvit.iiit.ac.in/research/projects/cvit-projects/the-iiit-5k-word-dataset>