A Study on Text Detection and Localization Techniques for Natural Scene Images

Salahuddin Unar
School of Computer Science and Technology, Faculty of
Electronic Information and Electrical Engineering,
Dalian University of Technology, Dalian 116024, China

Akhtar Hussain
Quaid-e-Awam University of Engineering, Science &
Technology, Nawabshah, Pakistan

Mohsin Shaikh
Quaid-e-Awam University of Engineering, Science &
Technology, Nawabshah, Pakistan

Kashif Hussain Memon
University College of Engineering & Technology
The Islamia University of Bahawalpur
Bahawalpur, Pakistan

Muhammad Adil Ansari
Quaid-e-Awam University of Engineering, Science &
Technology, Nawabshah, Pakistan

Zojan Memon
Quaid-e-Awam University of Engineering, Science &
Technology, Nawabshah, Pakistan

Abstract
In the current era of technology, information acquisition from the images become most important task due to the rapid development of data mining and machine learning. This paper introduces text detection from natural scene images. The paper provides analysis and compares different methods to detect and localize the text in natural scene images. The text exists in an image under varying properties such as text size, font, style, illumination, and orientation with complex background. We have addressed each state-of-the-art methods to resolve these issues. Next, experimental results and performance evaluations of several algorithms are presented. The results carried out under the various text detection approaches in form of precision and recall are given. Furthermore, different datasets are enlisted which are freely and publicly available, and performance protocols are also defined briefly.

Keywords
Text Detection; Text localization; Methodologies; Survey.

1. Introduction

This In our daily life, we observe several natural scene images containing very informative textual strings. The graphical images catch our attention and text contained in those images directs us to do some action or convey some important information. The text in images is informative not only for humans but its application varies in Robotics and Vehicles [1], impaired people assistance [2], analyzing the documents [3], guiding tourists and is a significant entity for retrieving and indexing purpose. Many researchers are paying attention to automatic text recognition in natural scene images because of potential applications in robotics, image retrieval, and intelligent transport system.

Detection of text from the natural images is a major task of computer vision. It attracts many researchers due to vast availability and easiness of handheld devices such as Mobile. However, the stated task is still the challenging one due to the varying properties of the text such as fonts, color, size, illumination and varying angle. The varying factors of text can be seen in Figure.1 (a-f).

The images can be divided into the three major groups including scene images, documental images, and born-digital images. Scene images consist of textual information and provide some instructions to follow just like captured images of posters, signboard, advertising banner. Documental images are the soft copy of the document and born-digital images are machine-born images contain the soft form of textual information. However, the textual information that exists in an image can be divided into two categories, i.e. the scene text and the artificial text. The primary type of text occurs in an image coincidently while the former type of text is presented through some computational techniques and plays a vital role to perceive and understand an image. So far, several methods based on binarization and edge detection have been proposed and applied to detect and recognize the textual information in scene images.

To detect and extract the text from the natural scene images, the process can be divided into the following three steps: (1) Text detection and localization, (2) Text extraction and enhancement, and (3) Recognizing the text (OCR). A systematic flowchart is given in Figure. 2. This paper focuses on the first step and explores several approaches developed for the detection and localization of the text.

Text detection and localization is the process of finding the position of the text in an image and creating bounding boxes surrounding the textual location by removing the complex background. For this purpose, preprocessing step is applied which converts the image into binary image and

Manuscript received January 5, 2018
Manuscript revised January 20, 2018

often enhanced by converting into the grayscale image. While in text extraction step, the textual components are segmented from the complex background. As the textual regions, generally are low resolution and can affect the result, so enhancement of textual regions is performed.

In this paper, we have analyzed, surveyed and summarized the different techniques and algorithms to detect and localize the text from the natural scene images so that the text image can be converted into the machine-readable format. There are many surveys which are based on old datasets, however, we have summarized and classified different methods and algorithms till 2016.

This paper primarily presents the survey of methods and techniques for localization and detection of text within images. More specifically the start-of-the-art mechanism for mentioned techniques is summarized and classified for better understanding. Furthermore, an introduction to Text detection and localization is presented and performance evaluations of several algorithms are briefly discussed. The paper has many interesting features in the context of text-mining of images.

The motivation behind this survey is to present different methods for text detection in an easy and simple way. The results are compared and datasets are given in tabular form for better and clear understanding to the readers.

2. Text detection and localization methodologies

The methodologies for detecting the text can be divided into the following four approaches: Connected-Component (CC) based, Texture-based, Edge-based, and Corner-based. Based on different approaches, the results are compared in Table 1. Each approach is discussed as follow:

2.1 Connected component-based approach

The connected component (cc) based approach considers the characters by segmenting the image, select the appropriate regions and locate the text from those detected regions. An image is subdivided into the smaller components and then these smaller components are grouped together to form the connected components until all regions of an image are determined [4] [5]. This approach segments textual components by color clustering or edge detection method. To recognize the textual components and classify them as textual regions, mostly a geometrical analysis is required as the CC-based approach is based on geometrical features. The approach has minimum computational complexity due to several segmented components. However, this approach is limited if the background color of the image and text is identical and if the image contains noisy and multicolored text. Another difficulty with this approach is that it cannot segment the textual components truly if the text position and scale is unidentified. In [9], Wang et al. proposed the CC-based method for recognizing character-like regions to detect and extract the text in the natural scene images. Alignment analysis is utilized to assure the block candidates once the connected components are extracted on multiple decompositions. The authors applied priority adaptive segmentation (PAS) to determine accurate character regions. Different heuristic processes like alignment properties and statistical features are used to verify the correct accuracy of segmented regions. Their method is robust for shooting conditions, colorful background and a wide range of textual fonts. However, the method is low effective under low contrast, irregular lighting, and multi-scale images.

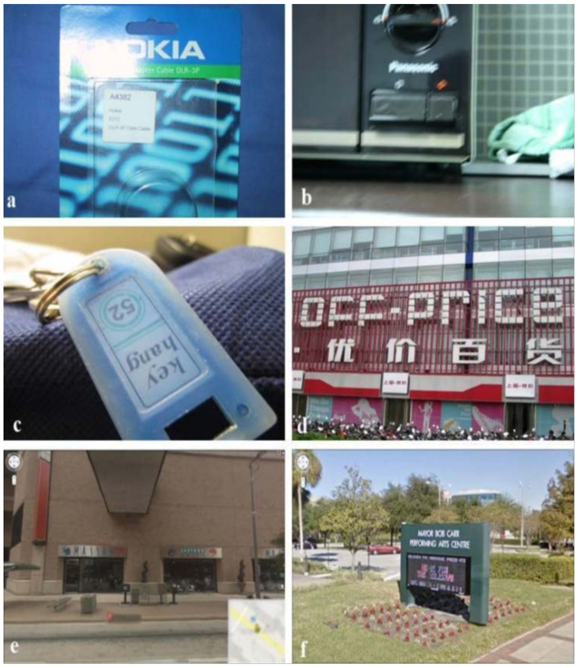


Fig. 1 Natural Scene Text Images under varying factors: fonts, color, size, illumination, varying angle. [6]–[8]

Felhi et al. [10] presented a new method for extracting text in the natural scene image. They proposed an essential descriptor that composes the strokes of text candidates and composes a spatial relation graph. Then a graph cut algorithm is employed to classify distinct nodes of the graph as text or non-text regions. They used a kernel SVM to improve the results further. However, the proposed method is not robust under alignment and varying orientation.

In [11], Jiang et al. proposed a learning-based method to detect and segment the text in natural scene images. An input image is partitioned into numerous connected-components by utilizing the Niblack clustering algorithm.

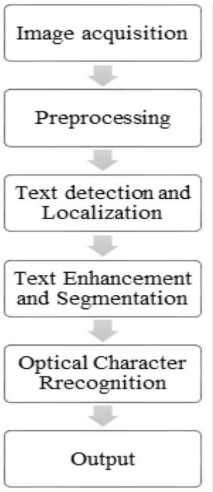


Image acquisition
Preprocessing
Text detection and Localization
Text Enhancement and Segmentation
Optical Character Rrecognition
Output

Fig. 2 the flowchart for text detection and recognition.

Once the image is decomposed then all the connected components (text CC and non-text CC) are verified for having textual features. The verification is done by using a two-stage classifier module as shown in Figure. 3. After running the second classifier, most of the non-text CCs are rejected and an SVM learning classifier verifies the remaining CCs further. After the classification, the positive textual CCs are fed into the binary image. The proposed method is robust for the different text size, color, angle, and font. However, the result is unsatisfactory for the text on a metal surface. Chucai et al. [12] given a new method for text detection with random variations in a complex scene image. The method is based on connected components grouping and image partition to find text regions embedded in the image. Firstly, they employed local gradient features and color uniformity to select candidate text characters in a partitioned image. Then performed character grouping on the bases of joint structural features of textual characters into text strings having minimum three character supporters in alignment. These features can be the character alignment, the difference of character size and neighboring features. The authors proposed two algorithms i.e. adjacent character grouping and text line grouping. The primary method estimates sibling groups of every character candidate as a separate string segment to merge cascaded sibling groups in text strings. The later method implements Hough transform for suitable text line that shows the orientation of a strong text string, as shown in Figure. 4. The proposed method lacks behind than the state-of-the-art methods for the horizontal orientations.

Ezaki et al. [13] proposed four CC-based character extraction methods. Each method's performance depends upon the character size. The most efficient extraction method follows the sequence: Sobel edge detection, Otsu local binarization, CC-based extraction and rule-based CC filtering. By combining all the candidate textual regions recommended by the each of the four methods, a high recall rate can be achieved. However, the method's performance is not enough for the practical applications. Yan et al. [14] have given a novel method for text detection and recognition according to which images divided into numerous layers by using color clustering. To determine the candidate text regions from the distinct layer, the authors employed connected component analysis, as shown in Figure. 5 and to determine either the candidate text region is text or non-text component, a cascaded Adaboost classifier is employed. The background noise efficiently reduced due to the monochrome color existence in each distinct layer and can improve the precision of text localization. An OCR application has utilized for recognizing the text regions determined by the cascaded classifier. The method is robust for complex background images.

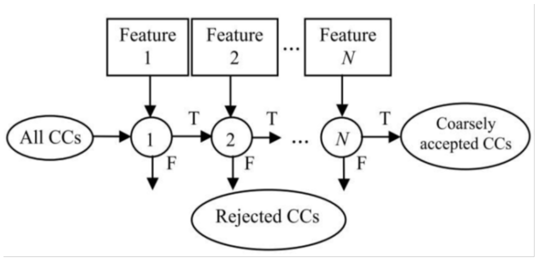
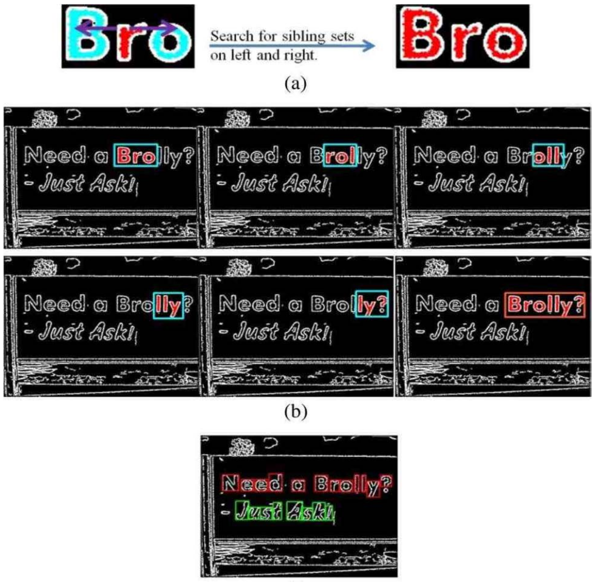


Fig. 3 Structure of cascaded Classifier. [11]



(a)

(b)

(c)

Fig. 4 (a) Sibling group of the connected-component “r”, Here “B” obtained from the left sibling group and “o” obtained from the right sibling group. (b) Merging the three sibling groups into a cascaded character group consistent with the text string “Brolly?” (c) Red and green are two detected cascaded character groups. [12]

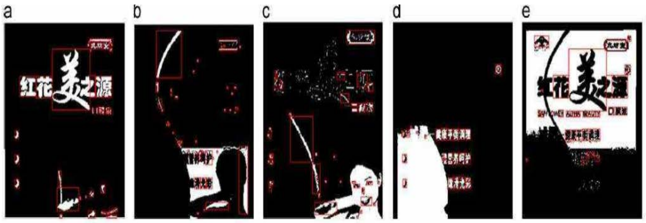


Fig. 5 Result of connected-components analysis. (a)Layer1 (b) Layer2, (c) Layer3, (d) Layer4 and (e) Layer5.

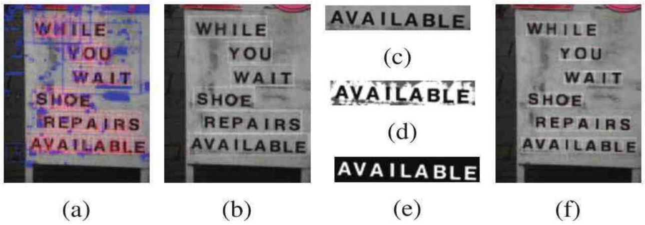


Fig. 6 An example of fine text localization. (a) Verifying patch (red= text, blue= non-text). (b) Grouping patch. (c) Text line image. (d) Local binarization. (e) Text components. (f) Text localization. [16]

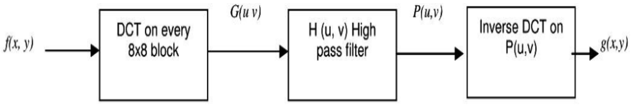


Fig. 7 High Pass Filter for Removing Background using DCT. [17]

2.2 Texture-based approach

The texture-based approach localizes the text boxes in an image by evaluating specific textual features over a window. The approach is used if the textual images have efficient textual characteristics that enable textual regions to distinguish them from the complex background. To extract textual characteristics from the image, this approach utilizes Support Vector Machine (SVM), Gaussian filtering, Fourier Transform, Wavelet decomposition, Discrete Cosine Transform, Neural Network, and Local Binary Pattern. This approach can efficiently detect the text either image is noisy or consisting complex background. Generally, a classifier (e.g. SVM) is used to classify the text and non-textual regions once the textual characteristics are extracted. The main advantage of this approach is speed and simplicity. However, this approach cannot be used for larger database due to the higher computational complexity, huge training dataset and its performance is also sensitive to text orientation and angle.

Hanif et al. [15] proposed a new scheme for detecting and localizing the text in grayscale images. The scheme utilized a small set of heterogonous features that are spatially shared to form a larger set of features. A localizer based on neural network learn the essential rules to localize the text. The authors utilized three distinct features including MDF (Mean Difference Feature), SD (standard deviation) and HoG (histogram of the oriented gradient) to extract the text segment on a block level. The proposed scheme is robust for various font sizes and styles in complex background images.

In [16], Pan et al. presented a new coarse-to-fine method for fast text localization by combining learning based region filtering and verification. A boosted region filter is employed to extract candidate text regions. By multi-orientation projection analysis, the text regions are segmented into the candidate text lines. A polynomial classifier with a boosted classifier is used for the fine verification of candidate text lines by discarding the non-texts candidates, as shown in Figure. 6. The polynomial classifier and boosted classifier both are fast for patch verification and region filtering respectively. The remaining candidate text patches are grouped into text lines based on the spatial relationship. Verification is done by using five distinct features including local binary pattern, the histogram of oriented gradients, discrete cosine transform, Gabor, and wavelets. The proposed method’s processing speed is high enough for practical applications.

Angadi et al. [17] proposed a texture based algorithm which uses a high pass filtering in DCT field to remove the constant background, as shown in Figure. 7. To classify the text regions, the feature vectors based on contrast and homogeneity are computed. The text features achieved at each 50x50 block of an input image and prospective text blocks are recognized using a discriminant function.

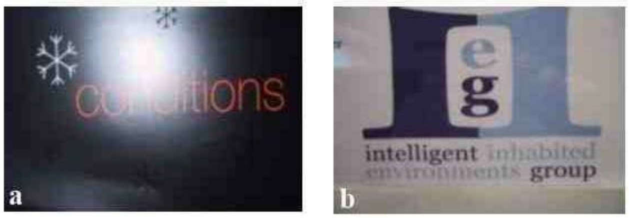


Fig. 8 (a) Illumination Variance (b) Text/Background Contrast Inconsistency. [19]

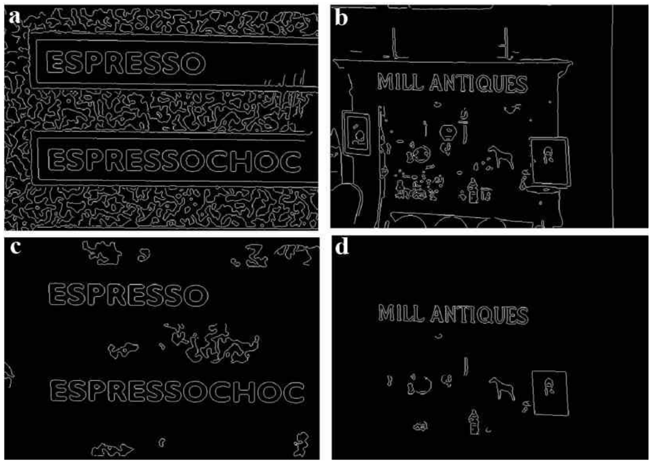


Fig. 9 An example of images processed by the edge filter. (a) & (b) are original edge maps, (c) & (d) are reserved edges after edge filter. [22]

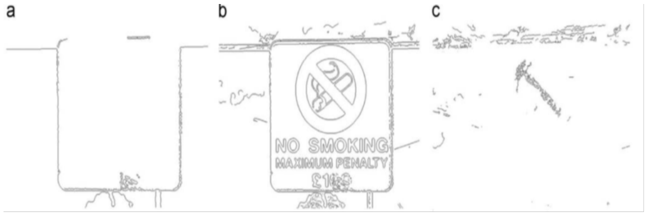


Fig. 10 An example of edges extracted from different channels: (a) Gray channel. (b) L channel. (c) Detail channel. [22]

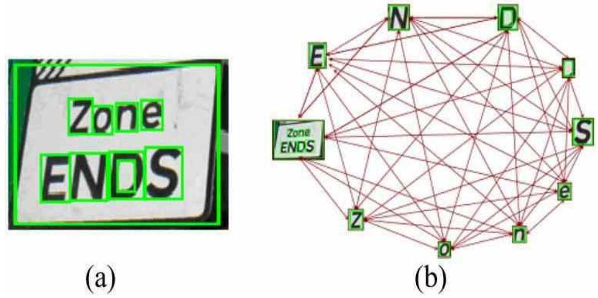


Fig. 11 (a) 10 candidate blocks (b) generated graph. [23]

The recognized blocks are combined to extract text regions. The proposed method is robust for varying text font, size, and alignment, and can detect nonlinear text regions. However, the authors mostly focused on localization of unbalanced text blocks.

In [18], Zhou et al. presented a novel method for multilingual text detection. Regardless of text language type, the method focuses to find the text regions in the scene image. The authors followed the same strategy as given in [15] with improved accuracy. Three distinct text features including MG (mean of a gradient), LBP (local binary patterns) and HOG (histogram of the oriented gradient) are selected to define multilingual text. After extracting the features from an image, a cascaded AdaBoost classifier is employed that combine distinct features. The proposed method is efficient for natural scene images containing multilingual text.

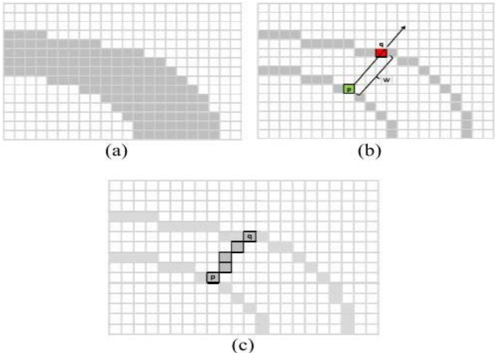


Fig. 12. Implementation of the SWT. (a) A classic stroke in which the pixel intensities of stroke are darker than the background pixels. (b) Here, p is a pixel on the border of stroke which is finding the track of the gradient at p leads to q, (c) Every highlighted pixel is assigned by the least of its current value and create the width of the stroke. [26].

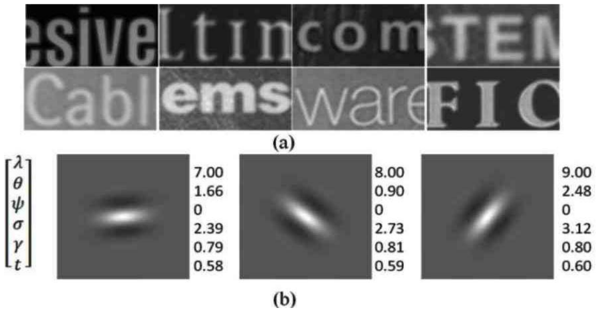


Fig. 13 (a) Text windows on positive training samples (b) SGWs with parameters on the right. [27]

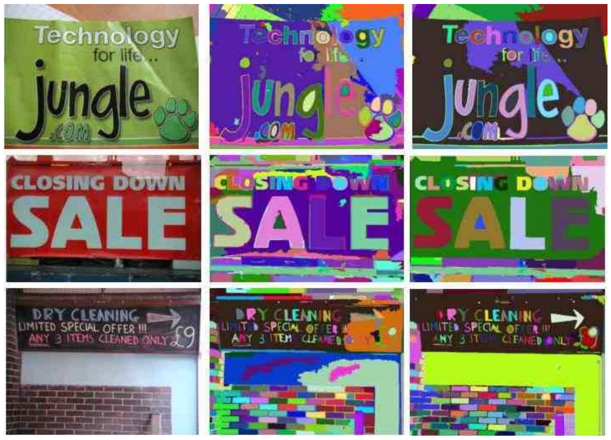


Fig. 14. Examples of text stroke segmentation. [29]

In [19], Ji et al. proposed an efficient method for text characterization based on LHBP (local haar binary pattern) to solve the problems of illumination variance and text background constant variance, as shown in Figure. 8. The framework combines the DCA-LHBP (Directional correlation analysis-Local haar binary pattern) with SVM-based post-classification. The authors’ extracted threshold restricted local binary pattern over the high-frequency coefficients of pyramid haar wavelets. While filtering steady illumination variances, it preserves unreliable text background contrasts. For finding candidate text regions, a directional correlation analysis (DCA) approach is proposed to filter non-directional LHBP regions. An SVM based classification using LHBP histogram is given to improve detection accuracy rate. The proposed method is robust for eliminating negative effects of illumination and text background contrast variance in natural scene images.

2.3 Distributed Edge based approach

The edge-based approach is simple and more efficient for extracting the text from natural scene images. It utilizes the structural and geometrical characteristics of the text. The edges are unique characteristics for detecting the text. Mostly an edge detector (e.g. Sobel, Canny) is used along with morphological operator in order to extract the text from the background and to remove non-textual regions. The better evaluation results can be obtaining from the images having solid edges but if the image having poor edges its performance is certainly reduced. So far, this approach is limited if the image has shadow and is blurred and edges are weak.

Gllavata et al. [20] presented an Edge-based algorithm that is an improvement of the method given in [21]. The algorithm detects, localize and extract horizontally aligned text having distinct fonts, size, and languages. The authors applied a local thresholding approach for arranging the histogram differences to improve the accuracy of text detection with complex background. The experiments are performed on a test set containing 175 images.

In [24], Ou et al. proposed an edge based method for text extraction. The method can detect and extract the text in the complex background image. The authors used three individual features of the text that may be utilized as unique characteristics to detect the text. The proposed method is effective for font size, color, alignment, and orientation and can be utilized in numerous applications such as license plate recognition, retrieving the document, robot navigation, and page segmentation.

Liu et al. [25] given an algorithm for detecting text in images and video frames, which include the three steps: edge detection, text candidate detection, and text refinement detection. Authors firstly applied edge detection to obtain four edge maps horizontally, vertically, up-left and up-right direction. Sobel edge detector is directly applied to the grayscale images and for color images, the color edge detector is applied. The color edge detector can be defined as:

$$I_\theta = \sqrt{\frac{1}{3}\sum_{i=1,2,3} I^2_{\theta C_i}} \qquad (1)$$

where, $I_{\theta Ci}$ is the edge map of i color channel in θ direction $i = 1,2,3$ and $\theta = 0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}$

Here, I_θ is the average edge map of all color channels in θ direction. Through this equation, the four directional edge maps can be obtained that show the edge density and edge strength in all of four directions. Then the features are extracted from the obtained four edge maps to label the texture characteristics of the text. To detect the initial text candidates, the K-means algorithm is applied. Finally, the text regions are recognized by the empirical rule analysis and processed through the project profile analysis. The evaluation is performed on a dataset of 100 images captured from web images, magazines, and real-life videos. The proposed algorithm is robust for font color, font size, multilingual and complex background.

Yu et al. [22] proposed an edge-based method for text detection and localization in natural scene images by combining edges, filtering edges and multi-channel processing. Edges are over segmented into the edge segments through the edge analysis in order to segment the text from the background and then the edge segments are recombined into candidate characters, as shown in Figure. 9. To filter out the background edges, an edge filter is utilized. The remaining candidate character edges link up to candidate text lines. The authors used two distinct classifiers to filter out the non-text lines. To confirm the recall, an adapted NMS (non-maximal suppress) and a multi-channel are used to avoid the duplication, as shown in Figure. 10. For improved classification accuracy, all the extracted edge-based and region-based features are kept into the feature pools and a linear SVM is employed to choose the appropriate features from the feature pools.

Zhang et al. [23] proposed an unsupervised text detection method based on HOG (histogram of the oriented gradient) and

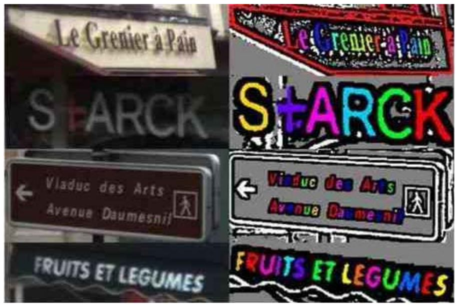


Fig. 15. TMMS segmentation on street-level image.

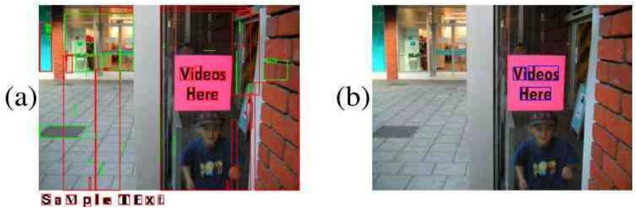


Fig. 16. Connected component’s clustering. (a) Recognizing Text and non-text CCs. (b) Merging text CCs to produce the final result. [32]

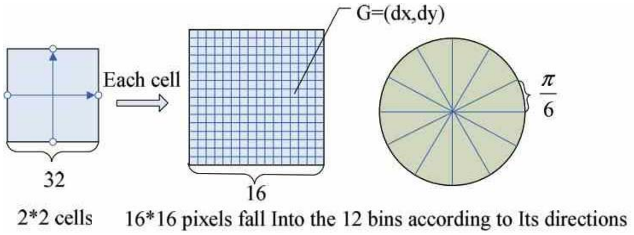

G=(dx,dy)
Each cell
32
16
π/6
2*2 cells
16*16 pixels fall Into the 12 bins according to Its directions

Fig. 17. An example of an illustration of the HOG feature. [33]

graph spectrum. To find the characteristics of text edges, the algorithm extracts the text edges from the image and localize candidate character blocks by applying HOG, as shown in Figure. 11. Then graph spectrum is employed that identifies the global connection between candidate character blocks and cluster candidate blocks in groups to create bounding boxed of text objects over the image. The proposed method is efficient for the text color, size, and orientation.

2.4 Strokes based approach

The stroke-based approach offers strong features for detecting the text in scene images. The text may demonstrate as a blend of stroke regions with an assortment of alignments and characteristics of text may be separated from mixtures and appropriations of the stroke segments. For stroke-based approaches, textual stroke competitors are separated by segmentation, proved by features extraction and characterization. This type of approach is easy to implement due to their perception and simplicity. However, the difficulty with this approach is complex background often sort the textual strokes difficult to segment and verify them.

Epshtein et al. [26] presented a novel image operator that locate the stroke width for each pixel. The operator is local and dependent on data due to which it effectively eliminates the need for multi-scale processing of windows. The authors defined the concept of a stroke and derive an algorithm to process it which create a new image feature, as shown in Figure. 12. After processing, the obtained feature is reliable and efficient for text detection. The algorithm combines the density estimation with the non-local scope. The proposed method is robust for multi-languages and varying fonts.

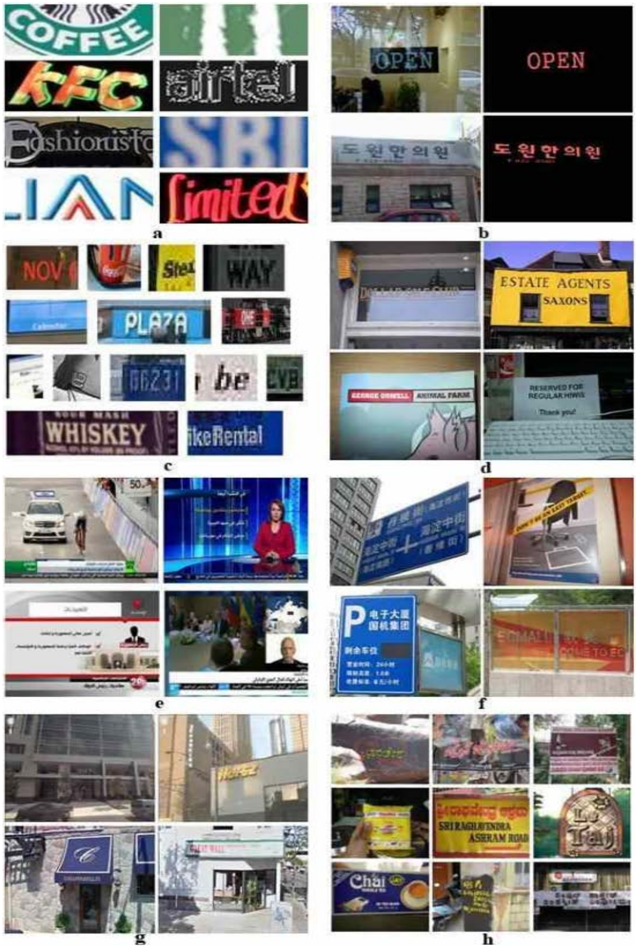


Fig. 18. Sample images of different datasets: (a) IIIT 5K-word [37] ;(b) KAIST [38]; (c) COCO-Text [39]; (d) ICDAR’13 [40]; (e) AcTiV [41]; (f) MSRA-TD500 [42]; (g) SVT [43]; (h) Char74K [44].

Yi et al. [27] proposed a novel approach to detect text regions in natural scene images based on stroke components and descriptive Gabor filter. According to which, the text characters and text string are created by stroke components as the basic units and Gabor filters are utilized to define and examine the stroke components in text characters or text strings. A set of Gabor filters is computed from the training set that can define the principle stroke components of text by the defined parameters. A K-means algorithm is employed which can cluster the descriptive Gabor filters. To provide a universal depiction of stroke components, the clustering centers are presented as SGWs (Stroke Gabor Words). Each SGW produce a couple of characteristics distribution for suitability measurement by evaluating positive and negative training samples. To analyze the suitability statistics, the Rayleigh model is used on positive samples and the Gaussian model is used on the negative samples respectively. To extract candidate image windows, heuristic layout analysis is applied first to a testing natural scene image. Then the principle SGWs are computed for each image window to present its principle stroke components, as shown in Figure. 13. The characteristics distribution created by SGWs are employed to classify the text and non-text windows. The proposed approach is robust for handling the varying text font, color, scale and complex background.

Karaoglu et al. [28] proposed an algorithm to detect and localize the text from indoor/outdoor images. The binarization method is based on the difference between the gamma correction and morphological reconstruction and used to extract the connected components of the image. By using a Random Forest classifier, these connected components are classified as the text or non-text. The Random Forest classifier handles many input variables and produces high classification accuracy. A novel merging algorithm is employed to localize the text regions. The proposed algorithm is robust for highlights, uneven illumination, overexposed, shadows and low contrast images. However, the proposed algorithm is time-consuming.

Pan et al. [29] proposed a hybrid method for text detection and localization in natural scene images by using stroke segmentation, verification and grouping. Firstly, the authors developed a scale-adaptive segmentation method to extract stroke candidates and then a CRF model based on local line fitting is designed with pairwise weight to verify the strokes, as shown in Figure. 14. For segmentation and verification accuracy, color-based text regions estimation is employed. The proposed method is robust for noisy and complex background images.

2.5 Other approaches:

However, because of several possible variations in textual contents, the above-given approaches often do not perform well and fail under several conditions like the complex background, non-horizontal text, non-uniform illumination and the variations of text size, font, orientation. To overcome such textual variations, the researchers have developed new hybrid methods that are a combination of above approaches.

In [30], Neumann et al. presented a method for text localization and recognition in real scene images. The method proceeds from a solid feed-forward pipeline and relocates it by a hypotheses verification framework directly processing multi-text line hypotheses. The synthetic fonts are used to train the algorithm discarding the unimportant acquisition and classify them real scene training data. The MSERs (Maximally stable extremal regions) algorithm is employed that describe accuracy for geometric and illumination conditions.

Fabrizio et al. [31] proposed a new hybrid approach to localize the text boxes that combines two approaches. A detection step is based on a CC-based approach and validation step to filter the text boxes is based on texture-based approach. The framework combines a hypothesis generation step to obtain potential text boxes and a hypothesis validation step to filter out incorrect detections respectively. The hypothesis generation step depends on a new efficient segmentation method that is based on a morphological operator. The authors introduced an efficient binarization TMMS (Toggle Mapping Morphological Segmentation) method to segment the images, as shown in Figure. 15. The regions are filtered and labeled by using the shape descriptors based on Pseudo Zernike moments, Fourier and an original polar descriptor which is invariant to the rotations. The three SVM classifiers are combined for the classification purpose in a late fusion scheme. The text box hypotheses are created from the characters detected in a group. The validation process is done by a global SVM classification of text boxes by using specified descriptors obtained from the HOG algorithm.

Mosleh et al. [32] proposed a text detection algorithm via the stroke width transform based on a feature vector created from connected components. A feature vector consists of distinct properties extracted from stroke width transform such as high contrast with the background, the variant directionality of gradient of text edges, and geometric characteristics of text components that exist in an image. Then component feature vectors are selected into a K-means clustering algorithm to isolate the potential text components from non-text ones. The selected text components are combined in a group and the remaining components are rejected, as shown in Figure. 16. The authors' presented a bandlet-based edge detector that is robust to select robust text edges and discard the noisy edges. The positions of the text-words are determined by utilizing the alignments of the selected text components. The proposed algorithm is robust for edge detection and text detection schemes.

In [33], Ma et al. presented text detection method based on edge detection and connected-component based approaches. A canny operator and an adaptive thresholding binarization method are employed for multi-scale edge detection. These detected edges are classified by SVM classifier integrating LBP, HOG, and several other distinct statistical features such as mean, standard deviation, entropy, energy, inertia, local homogeneity, and correlation. An example of the illustration of HOG features can be shown in Figure. 17. To filter out the non-text candidate regions and redetect the regions within the text candidates, a K-means clustering algorithm along with Otsu's algorithm is used.

Pan et al. [34] proposed a hybrid approach that detects and localize scene text by integrating region-based contents into the CC-based method. To segment candidate text components by local binarization, a text region detection is employed that evaluate the text exist confidence and scale evidence in an image pyramid. A CRF (conditional random field) model is utilized to consider unary component properties and binary text components relationship with supervised learning parameter. The textual components are combined into the text lines with a learning-based energy minimization algorithm.

Maruyama et al. [35] proposed a method to detect characters on signboards contain within natural scene images. A set of edge-based features (Haar wavelet and HOG) is utilized to determine character regions from non-character regions. Character detection is done by using the texture based features and character extraction is done by using the shape of the intensity distribution. The inconsistency between the detected regions and the normal distribution is obtained by skewness and kurtosis. The authors used these statistics with the texture-based features. A linear combination of stump classifiers is employed to detect character regions in the natural scene image. Finally, the feature components are selected for each stump and determination of coefficients of a linear combination is achieved by an AdaBoost algorithm.

3. Performance evaluation

3.1 Evaluation protocols

Most of the authors follow the formulae specified by ICDAR 2003 robust reading competitions [36], according to which precision and recall can be used to measure a retrieval system. The ICDAR evaluation protocols are most commonly used protocols for text detection. The evaluation of text detection and extraction from natural scene images based on the detection rate that consists of three parameters including precision p , recall, and f measure.

Precision p can be defined as the number of correct estimates divide by the total number of estimated:

$$p = \frac{c}{|E|}$$

The system that over-estimate the number of rectangles is categorized with a low precision value. The recall r is defined as the number of correct estimates divide by the total number of targets:

$$r = \frac{c}{|T|}$$

Hence, the system which under-estimate the number of rectangles is disciplined with a low recall value. Here, c represents the number of correct estimates, E represents the number of detected text, and T represents a ground-truth set of targets. However, the formulae defined above are unrealistic to evaluate the performance of text detection and extraction.

The authors defined the match M between two rectangles as the ratio between the area of intersection and the area of minimum bounding box consisting both the rectangles. The newly modified formulae for precision and recall can be given as:

$$p' = \frac{\Sigma_{r_e \in E} m(r_e, T)}{|E|}, \quad r' = \frac{\Sigma_{r_t \in \mathrm{T}} m(r_t, E)}{|T|}$$

The standard f measure is adapted to combine the precision p and recall r values into a single measure of detection quality, given as:

$$f = \frac{1}{\frac{\alpha}{p'} + \frac{(1-\alpha)}{r'}}$$

Hence, the f measure is the unit to indicate the performance of the algorithm and is harmonic mean of both precision and recall. The relative weights of precision and recall are controlled by α .

3.2 Datasets

The comprehensive collection of different datasets that are publicly accessible and commonly used for training and testing the text detection and localization are given in Table 2. The sample images from different public datasets as shown in Figure. 18.

4. Conclusion and future work

In this paper, we have analyzed and discussed different techniques for text detection from natural scene images. Different authors have followed different approaches and techniques to obtain satisfactory results. The comparative results of different approaches are highlighted in this paper. This analytical study gives the best way to understand the problem and possible direction to initiate the research in the field of text detection from natural scene images. The public datasets and performance measures are also described. Our future work will apply and implement connected-component based and edge based approach to detect the text from natural scene image.

Acknowledgment

This research is supported by the National Natural Science Foundation of China (Nos: 61672124, 61370145 and 61173183), the Password Theory Project of the 13th Five-Year Plan National Cryptography Development Fund (No: MMJJ20170203).

References

[1] N. K. Korghond and R. Safabakhsh, “AUT-UTP: Urban traffic panel detection and recognition dataset,” in 2016 24th Iranian Conference on Electrical Engineering (ICEE), pp. 1678–1682, 2016.

[2] S. P. F. Joan and S. Valli, “An enhanced text detection technique for the visually impaired to read text,” Inf. Syst. Front., pp. 1–18, 2016.

[3] R. Burduk, K. Jackowski, M. Kurzyński, M. Woźniak, and A. Żołnierek, “Text Detection in Document Images by Machine Learning Algorithms,” Adv. Intell. Syst. Comput., vol. 403, 2016.

[4] ICDAR’13 dataset, “http://rrc.cvc.uab.es/?ch=2&com=downloads,” 2013.

[5] The Street View Text Dataset, “http://vision.ucsd.edu/~kai/svt/,” 2011.

[6] MSRA Text Detection 500 Database (MSRA-TD500), “http://www.iapr-tc11.org/mediawiki/index.php/MSRA_Text_Detection_500_Database_(MSRA-TD500),” 2012.

[7] J. Ren-jie, F. Qi, L. Xu, and G. Wu, “Using Connected-Components Features to Detect and Segment Text,” J. Image Graph. 11, pp. 1653-1656, 2006.

[8] 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, 2000.

[9] Hao Wang and J. Kangas, “Character-like region verification for extracting text in scene images,” in Proceedings of Sixth International Conference on Document Analysis and Recognition, pp. 957–962, 2001.

[10] M. Felhi, N. Bonnier, and S. Tabbone, “A skeleton based descriptor for detecting text in real scene images,” Pattern Recognit. (ICPR), 2012 21st Int. Conf., no. Icpr, pp. 282–285, 2012.

[11] R. Jiang, F.-H. Qi, L. Xu, G. Wu, and K. Zhu, “A learning-based method to detect and segment text from scene images,” J. Zhejiang Univ. Sci. A, vol. 8, no. 4, pp. 568–574, 2007.

[12] C. Yi and Y. Tian, “Text string detection from natural scenes by structure-based partition and grouping.,” IEEE Trans. Image Process., vol. 20, no. 9, pp. 2594–2605, 2011.

[13] N. Ezaki, M. Bulacu, and L. Schomaker, “Text detection from natural scene images: towards a system for visually impaired persons,” Proc. 17th Int. Conf. Pattern Recognition, 2004. ICPR 2004., vol. 2, pp. 2–5, 2004.

[14] J. Yan and X. Gao, “Detection and recognition of text superimposed in images base on layered method,” Neurocomputing, vol. 134, pp. 3–14, 2014.

[15] [15] S. M. Hanif and L. Prevost, “Text detection and localization in complex scene images using constrained AdaBoost algorithm,” Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, pp. 1–5, 2009.

[16] Y. F. Pan, C. L. Liu, and X. Hou, “Fast scene text localization by learning-based filtering and verification,” Proc. - Int. Conf. Image Process. ICIP, pp. 2269–2272, 2010.

[17] S. A. Angadi, “A Texture Based Methodology for Text Region Extraction from Low Resolution Natural Scene Images,” International Journal of Image Processing (IJIP), no. 5, pp. 229–245, 2009.

[18] G. Zhou, Y. Liu, Q. Meng, and Y. Zhang, “Detecting multilingual text in natural scene,” Proc. 2011 1st Int. Symp. Access Spaces, ISAS 2011, pp. 116–120, 2011.

[19] R. Ji, P. Xu, H. Yao, Z. Zhang, X. Sun, and T. Liu, “Directional correlation analysis of local haar binary pattern for text detection,” 2008 IEEE Int. Conf. Multimed. Expo, ICME 2008 - Proc., pp. 885–888, 2008.

[20] J. Gllavata, R. Ewerth, and B. Freisleben, “Finding text in images,” ACM DL. pp. 539–542, 1997.

[21] J. Gllavata, R. Ewerth, and B. Freisleben, “A robust algorithm for text detection in images,” in 3rd International Symposium on Image and Signal Processing and Analysis, Proceedings of the ISPA 2003., vol. 2, pp. 611–616, 2003.

[22] W. Ou, J. Zhu, and C. Liu, “Text location in natural scene,” J. Chinese Inf. Process., vol. 5, p. 6, 2004.

[23] C. L. C. Liu, C. W. C. Wang, and R. D. R. Dai, “Text detection in images based on unsupervised classification of edge-based features,” Eighth Int. Conf. Doc. Anal. Recognit., pp. 0–4, 2005.

[24] Y. Feng, Y. Song, and Y. Zhang, “Scene text localization using extremal regions and Corner-HOG feature,” 2015 IEEE Int. Conf. Robot. Biomimetics, IEEE-ROBIO 2015, vol. 175, pp. 881–886, 2016.

[25] J. Zhang and R. Kasturi, "Text detection using edge gradient and graph spectrum," Proc. - Int. Conf. Pattern Recognit., pp. 3979–3982, 2010.

[26] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 2963–2970, 2010.

[27] C. Yi and Y. Tian, "Text Detection in Natural Scene Images by Stroke Gabor Words," 2011 Int. Conf. Doc. Anal. Recognit., pp. 177–181, 2011.

[28] S. Karaoglu, B. Fernando, and A. Trémeau, "A Novel Algorithm for Text Detection and Localization in Natural Scene Images," Int. Conf. Digit. Image Comput. Tech. Appl. 2010., pp. 635–642, 2010.

[29] Y. F. Pan, Y. Zhu, J. Sun, and S. Naoi, "Improving scene text detection by scale-adaptive segmentation and weighted CRF verification," Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, pp. 759–763, 2011.

[30] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6494 LNCS, no. PART 3, pp. 770–783, 2011.

[31] J. Fabrizio, B. Marcotegui, and M. Cord, "Text detection in street level images," Pattern Anal. Appl., vol. 16, no. 4, pp. 519–533, 2013.

[32] A. Mosleh, N. Bouguila, and a Ben Hamza, "Image Text Detection Using a Bandlet-Based Edge Detector and Stroke Width Transform," Proc. Br. Mach. Vis. Conf. 2013, pp. 1–12, 2013.

[33] L. Ma, C. Wang, and B. Xiao, "Text detection in natural images based on multi-scale edge detetion and classification," 2010 3rd Int. Congr. Image Signal Process., vol. 4, pp. 1961–1965, 2010.

[34] Y. F. Pan, X. Hou, and C. L. Liu, "A hybrid approach to detect and localize texts in natural scene images," IEEE Trans. Image Process., vol. 20, no. 3, pp. 800–813, 2011.

[35] M. Maruyama and T. Yamaguchi, "Extraction of characters on signboards in natural scene images by stump classifiers," Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, pp. 1365–1369, 2009.

[36] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "ICDAR 2003 robust reading competitions," Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, vol. 2003–Janua, pp. 682–687, 2003.

[37] S. M. Lucas, "ICDAR 2005 text locating competition results," Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, vol. 2005, pp. 80–84, 2005.

[38] D. Karatzas, S. R. Mestre, J. Mas, F. Nourbakhsh, and P. P. Roy, "ICDAR 2011 Robust Reading Competition - Challenge 1: Reading Text in Born-Digital Images (Web and Email)," 2011 Int. Conf. Doc. Anal. Recognit., pp. 1485–1490, 2011.

[39] D. Karatzas et al., "ICDAR 2013 robust reading competition," Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, pp. 1484–1493, 2013.

[40] D. Karatzas et al., "ICDAR 2015 competition on Robust Reading," Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, vol. 2015–Novem, pp. 1156–1160, 2015.

[41] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, "COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images," arXiv preprint arXiv:1601.07140, 2016.

[42] O. Zayene, J. Hennebert, S. Masmoudi Touj, R. Ingold, and N. Essoukri Ben Amara, "A dataset for Arabic text detection, tracking and recognition in news videos- AcTiV," Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, vol. 2015–Novem, pp. 996–1000, 2015.

[43] C. Yao, X. Zhang, X. Bai, W. Liu, and Y. Ma, "Detecting Texts of Arbitrary Orientations in Natural Images," Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE vol. 8, pp. 1–31, 2012.

[44] K. Wang, B. Babenko, and S. Belongie, "End-to-End Scene Text Recognition," Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, no. 4, 2011.

[45] Y. Netzer and T. Wang, "Reading digits in natural images with unsupervised feature learning," NIPS workshop on deep learning and unsupervised feature learning., vol. 2011, No. 2, pp. 1–9, 2011.

[46] A. R. Zamir and M. Shah, "Image Geo-localization Based on Multiple Nearest Neighbor Feature Matching Using Generalized Graphs," IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 8, pp. 1546–1558, 2014.

[47] T. E. de Campos, B. R. Babu, and M. Varma, "Character Recognition in Natural Images.," Visapp (2), pp. 273–280, 2009.

[48] S. Lee, M. S. Cho, K. Jung, and J. H. Kim, "Scene text extraction with edge constraint and text collinearity," Proc. - Int. Conf. Pattern Recognit., pp. 3983–3986, 2010.

[49] R. Nagy, A. Dicker, and K. Meyer-Wegener, "NEOCR: A conFigureurable dataset for natural image text recognition," International Workshop on Camera-Based Document Analysis and Recognition, vol. 6, pp. 150–163, 2012.

[50] A. Mishra, K. Alahari, and C. Jawahar, "Scene Text Recognition using Higher Order Language Priors," Procedings Br. Mach. Vis. Conf. 2012, p. 127.1-127.11, 2012.

Salahuddin Unar received his B.E degree in Computer Systems Engineering from Quaid-e-Awam University of Eng. Sc. & Tech., Nawabshah, Pakistan, and M.E degree in Computer & Information Engineering from Mehran University of Eng. & Tech., Jamshoro, Pakistan, in 2012 and 2015, respectively. Currently, he is pursuing Ph.D. degree in Computer Software and Theory from Dalian University of Technology, China. His research interest includes Image Processing, Image Retrieval, Text detection, Information retrieval, visual saliency, and pattern recognition.

Table 1: Result comparison of different methods

Author	Precision	Recall	F-measure	Approach	Determination	Dataset
Felhi et al. [10]	75	61	NA	Connect Component Based	Text detection in real scene image	ICDAR-2003, ICDAR-2011
Jiang et al.[11]	90.92	93.24	NA	Connect Component Based	Text detection and segmentation in natural scene images	NA
Chucai et al. [12]	71	62	62	Connect Component Based	Text string detection in natural scene images.	ICDAR-2003, OSTD
Ezaki et al. [13]	60	64	62	Connect Component Based	Text detection in natural Scene images	ICDAR-2003
Yan et al. [14]	83.4	82.6	NA	Connect Component Based	Text detection and recognition	ICDAR-2003
Hanif et al. [15]	28	13	NA	Texture Based	Text localization and classification	ICDAR 2003
Pan et al. [16]	66	70	68	Texture based	Text detection, feature extraction, classification	ICDAR 2003
Zhou et al. [18]	37	88	53	Texture Based	Text localization and classification, Multilingual scene text detection.	ICDAR 2003, Multilingual text dataset
Ji et al. [19]	59	79	68	Texture Based	Text illumination variance, Text-Background contrast variance.	ICDAR 2003
Gllavata et al. [20]	84.30	71.93	NA	Edge Based	Text detection, localization and extraction	MPEG 1 Video Frame News
Liu et al. [25]	78.3	81.5	NA	Edge Based	Edge detection, text candidate detection, text refinement detection.	100 Random Images
Yu et al [22]	84	65	73	Edge Based	Text localization, Feature pool, Edge analysis.	ICDAR 2011, SVT Database
Zhang et al. [23]	67	46	NA	Edge Based	Edge detection, Text edges extraction	ICDAR 2003
Epshtein et al. [26]	73	60	66	Stroke Based	Stroked width transforms	ICDAR 2003 ICDAR 2005
Yi et al. [27]	64	76	68	Stroke Based	Stroke components, Gabor filters	ICDAR 2003
Karaoglu et al. [28]	68	67	67	Stroke Based	Text localization, features extraction, forest classifier	ICDAR 2003 ICDAR 2005
Pan et al. [29]	68	67	67	Stroke Based	Text detection, stroke segmentation and verification.	ICDAR 2005
Neumann et al. [30]	59	55	57	Texture Based and Edge based	Text localization, MSER	ICDAR 2003 Chars75K
Fabrizio et al. [31]	63	50	NA	CC-Based and Texture-Based	Text detection and segmentation, TMMS, Toggle mapping	ICDAR 2005
Mosleh et al. [32]	76	66	71	CC-Based and Stroke-Based	Text localization, edge detection, feature vector.	ICDAR 2005
Ma et al.	67	72	NA	Edge Based and CC-Based	Edge detection, components analysis.	ICDAR 2003
Pan et al. [34]	67.4	69.7	68.5	Edge based and CC-Based	Text detection and localization, Conditional Random Field (CRF)	ICDAR 2005
Maruyama et al. [35]	85.6	88.7	NA	Edge Based and Texture Based	Character detection, Haar wavelet, HOG, intensity distribution	Random Signboard Images

Table 2: Public Datasets

Dataset	Properties	Source	Language	Web link
ICDAR 2003 [36]	Scene Images	Camera	English	http://algoval.essex.ac.uk/icdar/ Datasets.html Graphic
ICDAR 2005 [45]	Character Images	Camera	English	http://algoval.essex.ac.uk/data/icdar/ocr/digits/
ICDAR 2011 [46]	Scene Images, Graphic text Images	Camera	English	http://robustreading.opendfki.de/trac/
ICDAR 2013 [40]	Scene Images, Graphic text Images	Camera	English	http://dag.cvc.uab.es/icdar2013competition
ICDAR 2015 [47]	Scene Images, Graphic text Images, Web and email Images	Camera	English	http://rrc.cvc.uab.es/?ch=4&com=downloads
COCO-Text [39]	Scene Images, Character Images	Camera	English	https://vision.cornell.edu/se3/coco-text/
AcTiV [41]	Scene Images	Video frames	Arabic	http://tc11.cvc.uab.es/datasets/AcTiV_1
MSRA-TD500 [42]	Scene Images	Camera	English, Chinese	http://www.iapr-tc11.org/mediawiki/index.php/MSRA_Text_Detection_500_Database_(MSRA-TD500)
Street View Dataset (SVT)[43]	Scene Images	Video Frames	English	http://vision.ucsd.edu/~kai/svt/
Street View House Numbers (SVHN)[48]	Scene Images	Camera	English	http://ufldl.stanford.edu/housenumbers/
Google Street View [49]	Scene Images	Camera	English	http://crcv.ucf.edu/data/GMCP_Geolocalization/#Dataset
Char74k [44]	Character Images	Camera	English, Kannada	http://www.ee.surrey.ac.uk/CVSSP/demos/chars74k/
KAIST [38]	Scene Images	Camera	English, Korean	http://www.iapr-tc11.org/mediawiki/index.php/KAIST_Scene_Text_Database
NEOCR [50]	Scene Images	Camera	8 distinct languages	http://www.iapr-tc11.org/mediawiki/index.php/NEOCR:_Natural_Environment_OCR_Dataset
IIIT 5K-word [37]	Scene Images, Graphic text images.	Web based	English	http://cvit.iiit.ac.in/projects/SceneTextUnderstanding/IIIT5K.html