

Automated Latin Text Detection in Document Images and Natural Scene Images based on Connected Component Analysis

Muhammad Jaleed Khan¹, Naina Said², Aqsa Khan³, Naila Rehman³, Khurram Khurshid¹

¹*Vision Lab, Department of Electrical Engineering, Institute of Space Technology, Islamabad, Pakistan.*

²*Department of Computer Systems Engineering, University of Engineering & Technology, Peshawar, Pakistan.*

³*Faculty of Computer Science and Engineering, Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Pakistan.*

mjk093@gmail.com, nainasaid@uetpeshawar.edu.pk, aqsa.khan@giki.edu.pk, nailarehman74@gmail.com, khurram.khurshid@ist.edu.pk

Abstract—Robust and accurate detection of text in natural scene images and document images is a very challenging and common research problem. Over the past few decades, a variety of algorithms for text detection in images have been developed but there is still need for more robust and accurate text detection methods. In this work, we have proposed an accurate and robust text detection framework in which canny edge detection, maximally stable extremal regions and geometric filtering are employed in combination to efficiently collect and filter letter candidates in an image. Subsequently, individual letter patches are grouped to detect text sequences, which are then fragmented into isolated word patches. Finally, optical character recognition is employed to digitize the word patches. The proposed algorithm is tested on images representing different scenarios ranging from documents to natural scenes. Promising results have been reported which prove the accuracy and robustness of the proposed framework and encourage its practical implementation in real world applications.

Keywords—text detection; connected component; maximally stable extremal region; geometric checks; canny edge detector.

I. INTRODUCTION

Photos and videos constitute the major portion of the World Wide Web and they include noticeable text which is used to retrieve relevant content. Content based retrieval is a challenging problem that has received a significant amount of attention of researchers and experts in the domain of pattern recognition and computer vision. The key ingredients of any content based data retrieval system include text detection and character recognition. Many algorithms have been proposed to design better text detection and recognition systems over the past few decades. However, with everything going digital and the increasing complexity and amount of data on the web, the robustness and accuracy of text detection is becoming more challenging with time.

Contrary to the traditional text detection and recognition problems, where text is usually monotone on stationary and stable background, text detection in random images from the World Wide Web is comparatively far more complicated due to varying font, background, texture and lighting conditions [1]–[3]. Therefore, it is more challenging than ever to design a robust and accurate text detection framework which can address the aforementioned challenges. Recent work in pattern

recognition and document imaging includes the use of sophisticated solutions based on maximally stable extremal regions (MSERs) [4], [5], unsupervised learning [6], [7], hyperspectral image analysis [8], support vector machine [9], neural networks [1], stroke width transform [10], word spotting [11], [12] and deep learning [13], [14].

In this work, a novel text detection framework is proposed which is based on connected component analysis. Canny edges and MSER algorithms are employed for extraction of CCs, which are taken as letter candidates. CCs that are likely to be Latin characters are selected on the basis of their geometric properties. Afterwards, the selected objects are grouped to detect text sequences, which are then fragmented into isolated word patches. Optical character recognition is employed to digitize the word patches. The proposed algorithm is tested on images representing different scenes ranging from documents to natural scenes. Promising results have been reported which prove the accuracy and robustness of the proposed algorithm and encourage its practical implementation in real world scenarios.

Rest of the paper is structured as follows: Section II covers the background of the research problem addressed in this paper and related methods. The proposed algorithm is presented in Section III followed by experimental analysis in Section IV. Conclusion and future prospects of this research are summarized in Section V.

II. BACKGROUND

Text detection and recognition is a conventional problem that has been researched and constantly improved according to the increasing challenges in the images and videos on the web. There are some important terminologies associated with this research problem which are defined below;

- **Edge:** Edge is a group of points having strong gradient magnitude in an image.
- **Corner (or Point of Interest):** Corner is a group of points having a high level of curvature in the gradient in an image.
- **Region:** A region is a contiguous set of adjacent pixels.

- **Blob (or Region of Interest):** Blob is the area in which some properties (color, brightness, etc.) are invariant or slightly variant in an image, i.e. points in a blob are similar.
- **Boundary:** Boundary of a region is the group of pixels neighboring at least one pixel of that region but not a part of that region.
- **Extremal Region:** If all the pixels in a region have values greater than (or smaller than) that of the boundary, the region is called extremal region.
- **Maximally Stable Extremal Region (MSER):** An extremal region is termed as maximally stable when its variation w.r.t. a given threshold is minimal [4].

The modern methods used in the proposed system are reviewed in the following subsections.

A. Maximally Stable Extremal Region

MSERs are covariant regions in an image that do not vary noticeably through wide-ranging thresholds. The steps involved in extraction of MSERs in an image are as follows;

1) Thresholding:

Thresholding is applied in order to segment potential regions of interest. To avoid the effects of varying illumination conditions, local thresholding is usually preferred over global thresholding [6].

2) Extremal Regions:

Consider a sequence of threshold images with the threshold intensity sweeping from black to white images, we will find white spots in the images that appear and grow larger by merging. With increasing threshold, these white spots eventually merge and form the extremal regions representing blobs in the image.

3) MSER

Find the thresholds at which the extremal regions are “maximally stable”.

4) Region Descriptors

MSERs with ellipses are approximated as region descriptors. Ellipse serves as a region descriptor for the arbitrarily shaped MSERs.

B. Canny Edge Detection

Canny edge detection is one of the finest edge detectors introduced by John Canny [15]. A number of checks are performed are determined to improve its effectiveness. The checks include single response for single edge, localized edge points and low error rate. Canny edge detector comprises five steps as summarized below:

1) Gaussian Filtering

It is important to remove the noise and prevent false detection. The image is smoothened by applying a Gaussian filter. Performance of the detector depends on the filter size. Sensitivity of the detector is inversely proportional to the filter size, whereas the localization error is directly proportional to the filter size.

2) Intensity Gradient

Edges can point in various directions. Vertical, horizontal, and diagonal edges are detected in the image using four masks. The Sobel operator uses 3x3 convolutional masks to measure two dimensional spatial gradients. The direction of edge gradient is computed and rounded to one of four angles representing horizontal, vertical, and the diagonals, i.e. 0 , $\pi/4$, $\pi/2$ and $3\pi/4$.

3) Non-Maximum Suppression

In order to further sharpen the edges provided by gradients, all values in the mask except the local maximum are suppressed to zero. The local maximum represents the point change of intensity is the sharpest.

4) Double Thresholding

At this stage, the computed edges are a close approximation of the actual edge despite a few imperfections caused due to noise and color variation. To suppress these imperfections, edge pixels are classified as follows based on two thresholds, T_1 representing the higher threshold and T_2 representing the lower threshold:

- Strong Edge Pixels, if $p > T_1$ (preserved).
- Weak Edge Pixels, if $T_2 < p < T_1$ (conditionally preserved).
- Suppressed, if $p < T_2$.

5) Edge Tracking by Hysteresis

The extraction of weak edge pixels can be done from either noise/color variations or true edge. In order to achieve an accurate result, the weak edges caused from the noise should be removed. The weak edge is preserved only if there is at least one strong edge in its 8-neighbourhood.

III. PROPOSED TEXT DETECTION ALGORITHM

The proposed system is based on connected component analysis. Canny edges and MSER algorithms are employed for extraction of CCs, which are taken as letter candidates. CCs that are likely to be Latin characters are selected on the basis of their geometric properties. Subsequently, the selected objects are grouped to detect text sequences, which are then fragmented into isolated word patches. Optical character recognition is employed to digitize the word patches. The main steps involved in the proposed algorithm are presented in Figure 1 and an example is shown in Figure 2.

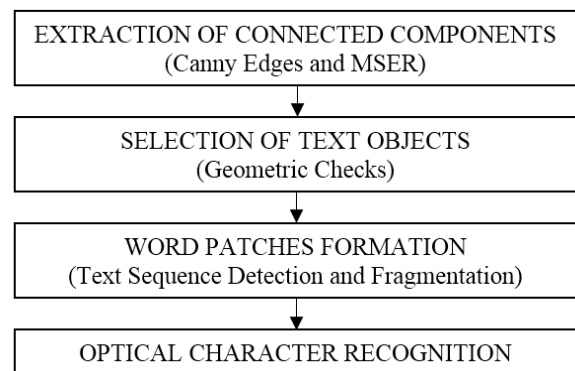


Fig. 1. Block diagram of the proposed text detection algorithm

A. Extraction of Connected Components

MSER is one of the best region detectors, but it is sensitive to image blur, due to which it fails to optimally detect small letters in low resolution images [16]. In order to overcome this limitation, MSERs and canny edges are computed and pixels in MSERs lying outside the canny boundaries are discarded. The CCs obtained after this step are taken as letter candidates.



(a)



(b)



(c)

Fig. 2. (a) An example image. (b) CC detection results. (c) Fine word patches formation after rejection of false positives.

B. Selection of Text Objects

After detecting letter candidates, some geometric properties of the CCs are checked and CCs that are not likely to be text objects are discarded.

- Objects which are either very large or very small are excluded.

- CCs with very large or small aspect ratio are excluded because the aspect ratio of most Latin letters is close to 1. A conservative threshold on the aspect ratio is chosen in order to keep CCs of elongated characters such as “l”, “I” and “i”.
- CCs with large number of holes are eliminated because they are no likely to be Latin letters.

C. Formation of Word Patches

Generally, text appears either in a linear form. The letter candidates are grouped pairwise to form text lines by using some ground rules.

- Letters in the same text line are assumed to exhibit same height.
- CCs that distant from one another are not grouped together.

As a result, the formation of text lines is based on collections of pair-wise grouped letter candidates. Within each cluster, centroids of the grouped CCs are calculated and a line intersecting with the largest number of CCs is taken as the text line. This whole process is repeated continuously unless all of CCs have been associated with a line or less than 3 CCs are present in a cluster. A line containing 3 or more text objects is declared as a text line.

Two additional validation steps are performed to filter out the improbable text lines. Numerous false positives are reported due to recurring objects, such as bricks and windows, in an image. A text line is rejected based on the following checks:

- If repetitive structures are observed in a major portion (like windows of an airplane)
- If numerous objects in a line have a very large solidity (letters are not solid, ‘A’, ‘O’...)

Finally, text lines are fragmented into isolated word patches by classifying the distances between letters as (i) character spacing, and (ii) word spacing.

D. Optical Character Recognition

The word patches are converted to digital form, specifically ASCII characters, using Optical character recognition [17]. The digitized text can then be used for content based retrieval or word processing.

IV. EXPERIMENTAL RESULTS

The proposed system was evaluated by conducting a series of experiments in a variety of scenarios ranging from document analysis to computer vision applications. Scanned documents were processed and digitized using the proposed algorithm. Natural images varying illumination conditions and complex background were captured. Some images of natural scenes and documents were taken from the internet as well.

The performance was remarkable on document images. All text was robustly and accurately detected by the proposed algorithm. In natural images, slight errors were reported, however the error rate was nearly negligible. The experimental results of the proposed algorithm on some examples of both types of images are presented in detail in Table I. The text detection results are shown along with recognized text. In the

recognized text column, wrongly recognized text is colored red for the sake of clarity.

Example no. 1-4 in Table I represent natural scenes relating to computer vision applications of the proposed algorithm. Example no. 1 is not much challenging as text is clear and easily detectable, thus perfect text detection and recognition is achieved. Example no. 2 represents a comparatively difficult scenario in which a sign board is captured from a large distance from a vehicle, which can relate to autonomous driving application of the proposed system. The text is accurately detected and recognized, however, small ASCII characters such as hyphen has been discarded by the algorithm as it does not match the geometric properties of characters. Example no. 3 represents a very challenging scenario with non-uniform illumination conditions and distorted text due to low resolution of the image. Due to use of local thresholding in the proposed algorithm, most of the text is accurately detected except small and distorted text in the end due to low resolution. In Example no 4, text is accurately detected and recognized but some of the very small text object in the last line on the board are wrongly recognized. It is hard for OCR to recognize distorted and very small text objects in low resolution images.

Example no. 5-6 in Table I represent document images relating to applications of the proposed system to document analysis and recognition. Example no. 5 is a historical document in which text is robustly detected and accurately recognized except a few letters are wrongly recognized. Example no. 6 is part of a scanned research article in which text is correctly detected and words are correctly recognized except some symbols. Some punctuation marks are not recognized in document images because small objects are discarded by the proposed algorithm.

The few limitations of the proposed algorithm noted in this section will be addressed in the future work. However, the promising results depict the efficacy and robustness of the proposed algorithm in complex scenarios which can be utilized in a variety of applications such as content based data retrieval, autonomous driving, robotics, web search, document analysis and recognition and word spotting. We will also extend the proposed algorithm for fine detection of oriental scripts such as Arabic [18] and Urdu [19].

V. CONCLUSION

A robust and efficient text detection algorithm based on connected component analysis is presented in this work. Canny edges and MSER algorithms are employed for extraction of CCs, which are taken as letter candidates. CCs that are likely to be Latin characters are selected on the basis of their geometric properties. Afterwards, the selected objects are grouped to detect text sequences, which are then fragmented into isolated word patches. Optical character recognition is employed to digitize the word patches. The proposed algorithm is tested on images representing different scenes ranging from documents to natural scenes.

Very promising experimental results have been achieved which depict the efficacy and robustness of the proposed algorithm in complex scenarios which can be utilized in a

variety of applications, such as robotic vision, content based data retrieval word spotting and document analysis and recognition.

REFERENCES

- [1] A. Yousaf, M. J. Khan, M. Imran, and K. Khurshid, "Benchmark dataset for offline handwritten character recognition," in *2017 13th International Conference on Emerging Technologies (ICET)*, 2017.
- [2] K. Khurshid, C. Faure, and N. Vincent, "Fusion of Word Spotting and Spatial Information for Figure Caption Retrieval in Historical Document Images," in *2009 10th International Conference on Document Analysis and Recognition*, 2009, pp. 266–270.
- [3] K. Khurshid, C. Faure, and N. Vincent, "Feature-based Word Spotting in Ancient Printed Documents," *undefined*, 2008.
- [4] H. Chen, S. S. Tsai, G. Schroth, D. M. Chen, R. Grzeszczuk, and B. Girod, "Robust text detection in natural images with edge-enhanced Maximally Stable Extremal Regions," in *2011 18th IEEE International Conference on Image Processing*, 2011, pp. 2609–2612.
- [5] A. C. Ozgen, M. Fasounaki, and H. K. Ekenel, "Text detection in natural and computer-generated images," in *2018 26th Signal Processing and Communications Applications Conference (SIU)*, 2018, pp. 1–4.
- [6] M. J. Khan, A. Yousaf, K. Khurshid, A. Abbas, and F. Shafait, "Automated Forgery Detection in Multispectral Document Images using Fuzzy Clustering," in *13th IAPR International Workshop on Document Analysis Systems*, 2018.
- [7] A. Abbas, K. Khurshid, and F. Shafait, "Towards Automated Ink Mismatch Detection in Hyperspectral Document Images," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017, pp. 1229–1236.
- [8] M. J. Khan, H. S. Khan, A. Yousaf, K. Khurshid, and A. Abbas, "Modern Trends in Hyperspectral Image Analysis: A Review," *IEEE Access*, vol. 6, no. 1, pp. 14118–14129, 2018.
- [9] C. S. Shin, K. I. Kim, M. H. Park, and H. J. Kim, "Support vector machine-based text detection in digital video," in *Neural Networks for Signal Processing X. Proceedings of the 2000 IEEE Signal Processing Society Workshop (Cat. No.00TH8501)*, vol. 2, pp. 634–641.
- [10] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2963–2970.
- [11] K. Khurshid, C. Faure, and N. Vincent, "Word spotting in historical printed documents using shape and sequence comparisons," *Pattern Recognit.*, vol. 45, no. 7, pp. 2598–2609, Jul. 2012.
- [12] K. Khurshid, C. Faure, and N. Vincent, "A Novel Approach for Word Spotting Using Merge-Split Edit Distance," Springer, Berlin, Heidelberg, 2009, pp. 213–220.
- [13] M. J. Khan, A. Yousaf, A. Abbas, and K. Khurshid, "Deep learning for automated forgery detection in hyperspectral document images," *J. Electron. Imaging*, vol. 27, no. 05, p. 1, Sep. 2018.
- [14] M. J. Khan, A. Yousaf, N. Javed, S. Nadeem, and K. Khurshid, "Automatic Target Detection in Satellite Images using Deep Learning," *J. Sp. Technol.*, vol. 7, no. 1, pp. 44–49, 2017.
- [15] J. Canny, "A Computational Approach to Edge Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.
- [16] C. Shi, C. Wang, B. Xiao, Y. Zhang, and S. Gao, "Scene text

detection using graph model built upon maximally stable extremal regions,” *Pattern Recognit. Lett.*, vol. 34, no. 2, pp. 107–116, Jan. 2013.

[17] D. Berchmans and S. S. Kumar, “Optical character recognition: An overview and an insight,” in *2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, 2014, pp. 1361–1365.


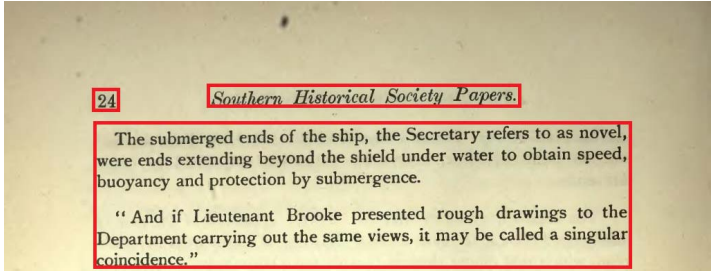
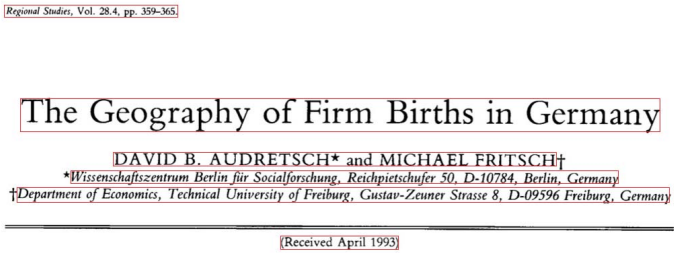
[18] U. Shahzad and K. Khurshid, “Oriental-script text detection and extraction in videos,” in *2017 1st International Workshop on*

Arabic Script Analysis and Recognition (ASAR), 2017, pp. 15–20.

[19] N. Javed, S. Shabbir, I. Siddiqi, and K. Khurshid, “Classification of Urdu Ligatures Using Convolutional Neural Networks - A Novel Approach,” in *2017 International Conference on Frontiers of Information Technology (FIT)*, 2017, pp. 93–97.

Table I. Detailed Experimental Results

S. No.	Category	Image with Text Detection Results	Recognized Text
1.	Natural Images		PLEASE TAKE NOTHING BUT PICTURES LEAVE NOTHING BUT FOOT PRINTS
2.			Rawat / Airport Sector I 8 H 8
3.			AlliedBank ATM Non Stop Banking Cards Accepted VISA

4.			<p>GOLRA SHARIF</p> <p>JN</p> <p>ME?N SEA LIVEL I994 54</p>
5.	Document Images		<p>24</p> <p>Southern Histovical Society Papers.</p> <p>The submerged ends of the ship, the Secretary refers to a novel, were ends extending beyond the shield under water to obtain speed,</p> <p>buoyancy and protection by submergence.</p> <p>“And if Lietenant Brooke presented rough drawings to the Department carrying out the same views, it may be called a singlar coincidence.”</p>
6.			<p>Regional Studies Vol 28 4 pp 359 365</p> <p>The Geography of Firm Births in Germany</p> <p>DAVID B AUDRETSCH and MICHAEL FRITSCH</p> <p>Wissenschaftszentrum Berlin f0r Socialforschung Reichpietschufer 50 D 10784 Germany</p> <p>Department of Economics Technical University of Freiburg Freiburg Gustav Zeuner Strasse 8 D 09596 Germany</p> <p>Received April 1993</p>