Automated Recognition of Text in Images: A Survey

Kanika Wadhawan
M.Tech Scholar
Christ University, Faculty of Engineering
Bengaluru, India

E. Gajendran, PhD
Research Guide
Christ University, Faculty of Engineering
Bengaluru, India

ABSTRACT

OCR (Optical Character Recognition) System works in the domain of Natural Language Processing and Image Processing. This is used to convert all the text information that is present in image form, to text format. Text is one of the most influential inventions of Humanity. The fertile and precise information incorporated in text is very useful in a wide range of applications that are computer-vision based, and hence text detection and recognition in natural scenes (e.g.: traffic sign boards, license plate, Hoardings and videos etc.) have become important and active research topics in computer vision and document analysis. This survey paper presents a review of various state-of-the-art techniques proposed for different processes (i.e. detection, localization, extraction, etc.) of text information processing in Images. Literature review can further serve as a good reference for researchers in the areas of scene text detection and recognition. The aim is to introduce the researchers to the latest trends in this area and to serve as a resource for developers who wish to integrate such solutions into their own work.

Keywords

Text Detection, Text Localization, Text Recognition, OCR

1. INTRODUCTION

The property of Natural Scenes Text usually carries high level semantics which makes text present in images and videos an important source of information. Localization and reading of texts in natural images are highly difficult tasks. The major challenges in text detection and recognition can be roughly categorized into three types [6, 7]:

• **Diverse nature of scene text:** In contrary to characters in document images, which are usually of regular font, same color, consistent size and uniform placement, texts in natural scenes may be of solely different scales ,fonts, colors, and orientations, even in one particular scene.

• **Complexity of the background:** The backgrounds in natural scene images and videos are generally quite complex. Components like fences, signs, bricks and grasses are virtually indistinguishable from true text, and hence are causing confusions and errors.

• **Interfering Elements:** Various interfering factors, for instance, noise, blur, distortion, low resolution, non-homogenous illumination and partial occlusion, may be an obstacle give rise to failures in scene text detection and recognition.

Solving these problems will require the applications of computer vision and pattern recognition techniques.

Several excellent review papers [21–23] exist in the fields of image text detection and recognition. Nonetheless, these survey papers are somewhat outmoded, since they had been published about ten years ago and have missed indefinite important, influential works that have been proposed in recent years.

Many researchers view Optical Character Recognition (OCR) as a solved problem. Text detection and recognition in imagery possess many of the same obstacles as computer vision and pattern recognition problems driven by lower quality or degraded data. Optical character recognition engines, such as Tesseract OCR [9], ABBYY Fine Reader [10], MODI (Microsoft Office Document Imaging) [11], are tuned up to detect and recognize the writings of text in scanned document images. Particularly, the recognition accuracy of Tesseract OCR is up to 97% when it is applied to scanned document images. The motive of this paper is to make attempts to establish the base by providing a comprehensive literature review of text detection and recognition research. Hence we summarize the problems and sub-problems, review the applications, and therefore analyze the challenges.

These methods can be broadly divided into three categories:

1) Text Detection.
2) Text Localization.
3) End-to-End text Recognition
As demonstrated in Fig. 1.

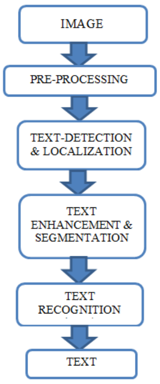
IMAGE
PRE-PROCESSING
TEXT-DETECTION & LOCALIZATION
TEXT ENHANCEMENT & SEGMENTATION
TEXT RECOGNITION
TEXT

Fig. 1 Flow Layout

2. RELATED WORK

Past many years, OCR systems have been an crucial application of machine learning, pattern recognition and computer vision. In the above research domains, prior works have mainly focused on systems that operate on scanned documents and on handwritten texts. Recently, some progress has been made in the field of text recognition in natural images and videos. A review of the new advancements achieved in the text recognition in multimedia documents is presented in [36]. Various issues related to the recognition problem have been identified, including text detection [9, 21, 23, 46], Enhancement of Text and linearization [3, 15, 24, 20, 48] (a pre-processing step that aims to improve recognition performance), character segmentation [2, 22, 31, 34], and its recognition [4, 17, 32] and the integration of linguistic knowledge [11, 27, 49].

Text detection approaches in [1-2] can be roughly classified into three groups:

- ❑ **Region-based approaches:** The attempt to use similarity criterions of text, such as color, size, stroke width, edge and gradient information, connected-components, to gather pixels.
- ❑ **Texture based approaches:** Texture-based methods use the observation that text in images has distinct textural properties so as to distinguish them from the background. The techniques based on Gabor filters, Wavelet, FFT, spatial variance, and many more can be used to detect the textural properties of a text region in an image.
- ❑ **Hybrid approaches:** This approach takes advantage of both region-based approaches which may closely cover text regions and texture-based approaches which may estimate coarse text location in scenes. E.g. Zhong et al. [6] fused the connected component (CC)-based approach with the texture-based approach.

Table 1: Text Detection Methods

METHOD	RULE	PROS	CONS
Using Boundary Features [1,3]	The boundary of license plate is rectangular	Simplest, fast and straightforward.	Hardly be applied to complex image since they are too sensitive to unwanted edges.
Using Color Features [2,6,5]	Specific color on license plate.	Be able to detect inclined and deformed license plates.	RGB is limited to illumination condition, HLS is sensitive to noise.
Using Texture Features [8,4]	Frequent color transition on license plate.	Be able to detect even if the boundary is deformed.	Computationally complex when there are many edges.
Using Character Features [7,9]	There must be characters on the license plate.	Robust to rotation.	Time consuming (processing all binary objects), produce detection errors when other text in image.

Table 2: Segmentation Methods

METHOD	PROS	CONS
Using Pixel Connectivity[1,2]	Simple and straightforward, robust to the license plate rotation.	Fails to extract all the characters when there are joined or broken.
Using Character Contours[4,9,3]	Can get exact character boundaries.	Sow & may generate incomplete or distorted contour.
Using Combined Features[1,3]	More reliable.	Computationally complex.

Table 3: Character Recognition Methods

METHOD	PROS	CONS
Using Pixel Connectivity[1,2]	Simple and straightforward, robust to the license plate rotation.	Fails to extract all the characters when there are joined or broken.
Using Character Contours[4,9,3]	Can get exact character boundaries.	Sow & may generate incomplete or distorted contour.
Using Combined Features[1,3]	More reliable.	Computationally complex.

3. METHODOLOGIES

In this section, we analyze two most commonly used methodologies in the complete text detection and recognition systems: stepwise and integrated. Stepwise methodologies have sorted out detection and recognition modules, and use a feeder pipeline to detect, segment and hence recognize text regions. Integrated methodologies, in contrast, have the aim to recognize words where the detection and recognition operations share some information with character classification and/or further use joint optimization strategies. Stepwise methodologies employ a coarse-to-fine strategy, which firstly localizes the text candidates, and so verifies, segments, and recognizes them. One of the attractive features is that the most of the background is filtered in the coarse localization step, which immensely reduces the computational cost, and eventually guarantees computational efficiency. The second attractive feature is that it processes oriented text as the text orientations and are approximated in the localization step. Language independent features or multilingual OCR modules in [12], [27], [34], processes multilingual text. The disadvantages are dual. The first increases the complexity when integrating different techniques from all steps. Second is the difficulty in optimizing parameters for all steps, which could introduce error accumulation. By counterpoint, the goal of integrated methodologies is to identify the specific words in imagery with character and language models. Integrated methodologies might avoid the challenging segmentation step or optimizing it with character and word recognition, which makes it a little less sensitive to complex backgrounds and low resolution text. The disadvantage lies in the multi-class character classification procedure which is computationally expensive in considering a large character class number and a large amount of candidate windows. Further, the increase of word class number might significantly decrease the detection and recognition performance; hence the generality is often limited to a small lexicon of words.

4. FUNDAMENTAL SUB-PROBLEMS

Under this section, sub-problems including text localization, segmentation, and recognition are discussed. Every approach is viewed with respect to its contribution. The techniques which make multiple contributions are analyzed with respect to each techniques contribution.

4.1 Text Detection and Localization

On the basis of features used, detection of text, localization Techniques can be categorized in two categories [3, 4], viz, *Region-based* and *Texture based.* The *Region based* technique works in a bottom-up fashion, by dividing the frame into sub regions and then merging the likely text regions to form bounding boxes for the region containing text. In the *Region based* approaches connected components, color, and edge features are commonly used. The texture properties of the text are used to differentiate between the text and background are used in *Texture based* methods. Techniques such as Wavelet transform, Gabor filters, Discrete Fourier transform, and machine learning, etc. are often used in *Texture-based* techniques. Brief overview of the techniques proposed in each of the mentioned categories is present in sub-section A and B.

4.1.1 Region based methods

A stroke width similarity based approach for text detection was projected by Din*ette al.* [8]. Some local adaptive threshold was used to nullify the background while preserving the text. Dilation, one of the morphological operations was applied for text localization. For refinement of the text location, use of multi-frame refinement method was done. In fact stroke filter based methods [9, 10, 11], and Stroke Width transform [12] were in use by many researchers for detection and localization of text. Jung *et al.* [9] used stroke filter method for text segmentation, counting the intrinsic features of the text. On the basis of stroke filter response and text polarity, growing of local region was used for the segmentation of the text. An OCR feeder score was used to improve the accuracy of text segmentation. Jung *et al.* [11] discussed the stroke filter technique and its application in text localization in video frames in detail. Classifiers such as SVM classifier is used for the verification of the text candidates. Based on the score of verification and color distribution, the line refinement of text is done. Li *et al*. [10] uses stroke filter approach to calculate the stroke map. They made use of two SVM classifiers to obtain the rough text regions and hence to verify the text lines candidates. Localization of text was achieved by projection profile. The second SVM used to verification of localized text lines. Shiva kumara *et al.* [13] proposed an edge based technique to detect text in images which is present in the horizontal direction. The frame was segmented into 16 non-parallel blocks. Mean-median filter and edge analysis were used to discover the candidate text blocks. Use of block growing method, to get the complete text block. Eventually, based on the vertical and horizontal bar features, the true text regions were detected. In Shiva kumara*et.al* [14], filters and edge analysis were used for detection of initial text. The straight and cursive edge features were used for elimination of all false positives.

Park *et al.* [21] used vertical and horizontal projection visibility to detect Korean text in out-of-door signboards. Shiva kumara *et al*. [22] proposed the classification of low and high contract images for detection of text. They examined the number of edges that could be found using sober and canny edge detector algorithm for images with low and high contrast, to figure the heuristic rules for text classification. A Self-Organizing Map (SOM) neural network approach for text detection in video frames was because of Yu *et al.* [20]. Pan *et al.* [36] introduced a hybrid method for text detection and localization making use of stroke segmentation, verification, and at last grouping.

4.1.2 Texture based methods

Wavelet transforms and its variants have become very famous among researchers for analysis of texture. Much of the recent work on texture based text detection and localization technique are based on the wavelet transform [38, 39, 40, 41, 42]. Various other methods such as the Gabor filter [43, 44], DCT [51], spatial analysis [46], Haar wavelet [27], Fourier [48], Laplacian [20] and many more, were also in great use by researchers in the past.

Combination of wavelet features and an SVM classifier were used in [38, 40, 27]. Ye *et al*. [38] made use of the 2D wavelet coefficients for calculating histogram wavelet coefficients of all pixel values. SVM together with the RBF kernel was used to classify text and non-text. They further introduced an OCR feedback procedure for locating the final text lines. Ji *et al.* Phan *et al.* [50] used the very same Laplacian approach as in [20] for identification of text candidates, but also used Connected Component analysis to form simple CCs. Using the straightness and edge density features, the text blocks were settled. Fourier features in RGB space were used for detection of text in video frames by Shiva kumara*et.al.* [48].

All texture intensities were used to affirm horizontal and vertical text. Horizontal and Vertical projection profiles were used for localization of text.

4.2 Extraction, Binarization, and Enhancement

In the recent past, not much work has been done towards text extraction, binarization and enhancement. They mainly aim towards the extraction of the single characters from the detected and localized blocks of text, for the OCR System. Quite wide range of binarization techniques have been used in past few decades to get a two tone image. In order to improve OCR accuracy, enhancement of the extracted single characters is needed, A K-means clustering and SVM based technique for binarization was discussed by Wakahara *et al.* [38], which is a four step technique. HSI color space was helpful in distinguishing characters against the backgrounds. SVM was used to determine whether the image is character or non-character images. Alike Characters estimations are used to achieve optimized binarized result. Ntrirogiannis *et al.* [39] used the lower and upper baseline of the text, convex hull analysis and stroke width for binarization of the text in video frames. Binarization technique introduced by Zhou *et al.* [36] used the contour of the text together with local thresholding to calculate the inner side of the contour. The contour is then filled up to form into the characters. Mishra *et al.* [37] presented an MRF based approach using binarization of natural image text. Character recognition accuracy of 95% was reported. A recognition scheme for Korean characters present in the out-of-door signboards was introduced by Park *et al.* [21]. The System made use of a minimum distance classifier together with a shape based statistical feature, for recognition of character. An Arabic video for text recognition was discussed by Halima *et al.* [23]. The feature used for recognition included some projection features, occlusion features, transition features, number of components in the character and dots location. A *k*-nearest neighborhood classifier was used for classifying, and optimized results were obtained for k=10. Iwamura *et al.* [66] proposed a non-learning based approach for camera captured characters. It tries to look-for the most alike example of an input character. Saidane and Gracia [31] made use of a convolutional neural network for character recognition and gathered an average recognition accuracy of 84.53% from ICDAR 2003 dataset. Features which were used included oriented, corners edges, end points that were achieved directly from the three color channels. Saidane *et al.* [65] proposed a graph based technique called image Text Recognition Graph (iTRG) for color text recognition in images and videos. The graph consisted of five modules, such as, graph connection builder, character recognition, text segmentation, and optimal path search module and graph weight calculator. ICDAR 2003 data set was used for performance evaluation. Coates *et al.* [68] proposed a scheme which was based on unsupervised feature learning. A combination of linear SVM and K-means clustering, gave 81.7% accuracy .The dataset ICDAR 2003 was used for testing.

5. EVALUATION PROTOCOLS

5.1 Evaluation protocols for text detection algorithms

In image text detection, there are three most important metrics in performance assessment: recall, precision and F-measure. Precision counts the ratio amongst the true positives and all detections, whereas recall measures the ratio of the true positives and all the true texts that has to be detected. F-measure, an overall, unit indicator of algorithm performance and is the harmonic mean of recall and precision.

The match *m* within two rectangles is the ratio of the area of convergence of the rectangles and that of the minimum bounding rectangle that contains both. The set of rectangles judged by each algorithm are named *estimates* and the set of

Ground truth rectangles in the ICDAR dataset are named *targets*. For each rectangle, the match with the highest value is found. Therefore, the best match for a rectangle *r* in a set of rectangles *R* can be defined as:

$$m(r;R) = max\{m(r,r\_|r\_ \in R\}$$

Then, according to the definitions of precision and recall are:

$$precision = \frac{\sum_{re\in E} m(re;T)}{|E|}$$

$$recall = \frac{\sum_{rt\in T} m(rt;E)}{|T|}$$

Where *E* and *T* are said as the sets of ground truth rectangles and estimated rectangles, respectively. F-measure *f* is said as a combination of the two above measures, *recall* and *precision.* The relative weights of recall and precision are tackled by a parameter α, which is generally set to 0.5 to give equal weightage to precision and recall:

$$f = \frac{1}{\frac{\alpha}{precision} + \frac{1-\alpha}{recall}}$$

The protocol implemented by Wolf et al. [84] considers three matching cases: one-to-one case, one-to-many case and many-to-many case. Recall and Precision are defined as:

$$precision\ (G,D,tr,tp) = \frac{\sum_j MatchD(Dj,G,tr,tp)}{|D|}$$

$$recall = \frac{\sum_i MatchG(Gi,D,tr,tp)}{|G|}$$

$$MatchD(Dj,G,tr,tp) = \begin{cases} 1, & if\ one - to - one\ match: \\ 0, & if\ no\ match; \\ fsc(k), & if\ many\ (\rightarrow k) matches \end{cases}$$

$$MatchG(Gi,D,tr,tp) = \begin{cases} 1, & if\ one - to - one\ match: \\ 0, & if\ no\ match; \\ fsc(k), & if\ many\ (\rightarrow k) matches \end{cases}$$

Object detection task [91], in the protocol with Refs. [6, 12] detections can be considered true or false positives on the basis of overlap ratio between the estimated minimum area rectangles and the ground truth rectangles. E.g. If the included angle for the estimated rectangle and the ground truth rectangle is less than π/8 and moreover their overlap ratio exceeds 0.5, the calculated rectangle is considered as a correct detection. Multiple or more than one detections of the same text line are considered as false positives. The definitions of recall and precision are:

$$precision = \frac{|TP|}{|E|} \qquad recall = \frac{|TP|}{|T|}$$

5.2 Evaluation protocols for text recognition algorithms

In image text recognition, the functionality of an algorithm can be measured either by character level recognition rate or by word level recognition rate. We have discussed a comprehensive literature survey on image text detection and recognition

6. ACKNOWLEDGMENTS

Heartiest thanks to the expert Guide, Mr. E. Gajendran, Assistant Professor, Christ University, Faculty of Engineering, Bengaluru, who has contributed towards development of the template and the research.

7. REFERENCES

[1] Casey, R., Lecolinet, E.: A survey of methods and strategies in character segmentation. IEEE Trans. Pattern Anal. Mach. Intell.18 (7), 690–706 (2002)

[2] Chen, D., Odobez, J., Bourlard, H.: Text detection and recognition in images and video frames. Pattern Recogn.37 (3), 595–608 (2004)

[3] Chen, T., Ghosh, D., Ranganath, S. : Video-text extraction and recognition. In: IEEE Region 10 Conference, TENCON'04, vol.1, pp. 319–322 (2005)

[4] K. Jung, K. I. Kim, and Anil. K. Jain, “Text information extraction in images and video: a survey”, Pattern Recognition, vol. 37, 2004, pp.977-997.

[5] J. Zhang and R. Kasturi, “Extraction of Text Objects in Video Documents: Recent Progress”, DAS, 2008, pp.5-17.

[6] Yao C, Zhang X, Bai X, Liu W, Tu Z. Rotation-invariant features for multi-oriented text detection in natural images. PloS one, 2013, 8(8):e70173

[7] Yao C, Bai X, Shi B, LiuW. Strokelets: A learned multi-scale representation for scene text recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2014, 4042–4049

[8] Delakis, M., Garcia, C.: Text detection with convolutional neural networks. In: International Conference on Computer Vision Theory and Applications, vol. 2, pp. 290–294 (2008)

[9] ABBYY Fine Reader.http://finereader.abbyy.com/

[10] Microsoft Office Document Imaging. http://en.wikipe_dia.org/wiki/Microsoft_Office_Document_Imaging

[11] M. Cai, J. Song and M.R. Lyu, ”A New Approach for Video Text Detection,” in Proc. IEEE Int’l Conf. Image Processing, pp.117-120, 2002.

[12] Yao C, Bai X, Liu W, Ma Y, Tu Z. Detecting texts of arbitrary orientations in natural images. In: Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. 2012, 1083–1090

[13] C. Jung, Q. Liu, and J. Kim, “A new approach for text segmentation using a stroke filter”, Signal Processing, vol.88, 2008, pp.1907-1916.

[14] X. Li, W. Wang, S. Jiang, Q. Huang, and Wen Gao, “Fast and Effective text detection”, ICIP, 2008, pp.969-972.

[15] C. Jung, Q. Liu, and J. Kim, “A stroke filter and its application to text localization”, Pattern Recognition Letters, vol.30, 2009, pp.114-122.

[16] B. Epshtien, E. Ofek, Y. Wexler, “Detecting text in natural scenes with Stroke Width Transform”, CVPR, 2010, pp. 2963 - 2970.

[17] P. Shiva kumara, W. Huang, C. L. Tan, “An Efficient Edge Based Technique for Text Detection in Video Frames”, DAS, 2008, pp.307-314.

[18] Hua, X., Yin, P., Zhang, H.: Efficient video text recognition using multiple frame integration. In: International Conference on Image Processing, vol. 2, pp. 397–400 (2002)

[19] Li, H., Doermann, D., Kia, O.: Automatic text detection and tracking in digital video. IEEE Trans. Image Process.9 (1), 147–156(2000)

[20] Yi C, Tian Y L. Text string detection from natural scenes by structure based partition and grouping. IEEE Transactions on Image Processing, 2011, 20(9): 2594–2605

[21] J. Park, G. Lee, E. Kim, J. Lim, S. Kim, H. Yang, M. Lee, and S.Hwang, “Automatic detection and recognition of Korean text in outdoor signboard Images”, Pattern Recognition Letters, vol.31,2010, pp.1728-1739.

[22] P. Shiva kumara, W. Huang, T. Q. Phan, C. L. Tan, “ Accurate videotext detection through classification of low and high contrast Images”, Pattern Recognition, vol.43, 2010, pp.2165-2185.

[23] P. Shiva kumara, A. Dutta, U. Pal, and C. L. Tan, “A New method for Handwritten scene text detection in video”, ICFHR, 2010, pp.387-392

[24] J. Yu and Y. Wang, “Apply SOM to Video Artificial text area detection”, Int. Conf. Internet computing for Science. And Engg.,2010, pp.137-141.

[25] X. Huang and H. Ma, “Automatic Detection and Localization of Natural scene text in Video”, ICPR, 2010, pp.3216-3219.

[26] Li H, Doermann D, Kia O. Automatic text detection and tracking in digital video. IEEE Transactions on Image Processing, 2000, 9(1):147–156

[27] M.R. Lyu, J. Song, and M. Cai, ”A Comprehensive Method for Multilingual Video Text Detection, Localization, and Extraction, ”IEEE Trans. Circuits System on Video Technology, vol. 15,no. 2, pp. 243-255, 2005.

[28] Yi, J., Peng, Y., Xiao, J.: Using multiple frame integration for the text recognition of video. In: International Conference on Document Analysis and Recognition, pp. 71–75 (2009)

[29] Yokobayashi, M., Wakahara, T.: Segmentation and recognition of characters in scene images using selective binarization in color space and GAT correlation. In: International Conference on Document Analysis and Recognition, pp. 167–171 (2005)

[30] Shivakumara P, Phan T Q, Tan C L. A laplacian approach to multioriented text detection in video. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(2): 412–419

[31] Z. Saidane and C. Gracia, “Automatic Scene Text Recognition using a Convolutional Neural Network”, CBDAR, 2007, pp.100-107.

[32] A. Ohkura, D. Deguchi, T. Takahashi, I. Ide, and H. Murasse, “Low resolution Character Recognition by Video-based Super-resolution”, ICDAR, 2009, pp.191-19.

[33] Z. Saidane, C. Garcia, and J. L. Dugelay, “The image Text Recognition Graph (iTRG)”, ICME, 2009, pp.266-269

[34] P. Shiva kumara, W. Huang and C.L. Tan, ”Efficient Video Text Detection using Edge Features,” in Proc. IEEE Int’l Conf. Pattern Recognition, pp. 1-4, 2008.

[35] Everingham M, Van Gool L, and Williams C K I, Winn J, Zisserman A. The pascal visual object classes (voc) challenge. International Journal of Computer Vision, 2010, 88(2): 303–338

[36] Z. Zhou, L. Li, C. L. Tan, “Edge based Binarization for video text images”, ICPR, 2010, pp.133-136.

[37] A. Mishra, K Alahari, and C. V. Jawahar, “An MRF Model for Binarization of Natural Scene Text”, ICDAR, 2011, pp-11-16.

[38] Toru Wakahara and Kohei Kita, “Binarization of Color Character Strings in Scene Images Using K-Means Clustering and Support Vector Machines”, ICDAR, 2011, pp.274-278.

[39] K. Ntirogiannis, B. Gatos, and I. Pratikakis “Binarization of Textual Content in Video Frames”, ICDAR, 2011, pp.673-677.