A COMPREHENSIVE STUDY ON TEXT INFORMATION EXTRACTION FROM NATURAL SCENE IMAGES

Anit V. Manjaly[1] and B. Shanmuga Priya[2]
School of Computer Science, CMS College of Science & Commerce, India
E-mail: [1]manjaly.anit@gmail.com, [2]shanmugapriyaa.b@gmail.com

Abstract

In Text Information Extraction (TIE) process, the text regions are localized and extracted from the images. It is an active research problem in computer vision applications. Diversity in text is due to the differences in size, style, orientation, alignment of text, low image contrast and complex backgrounds. The semantic information provided by an image can be used in different applications such as content based image retrieval, sign board identification etc. Text information extraction comprises of text image classification, text detection, localization, segmentation, enhancement and recognition. This paper contains a quick review on various text localization methods for localizing texts from natural scene images.

Keywords:

Natural Scene Image, Text Information Extraction (TIE), Text Detection, Text Localization, Segmentation, Extraction, Enhancement, Recognition

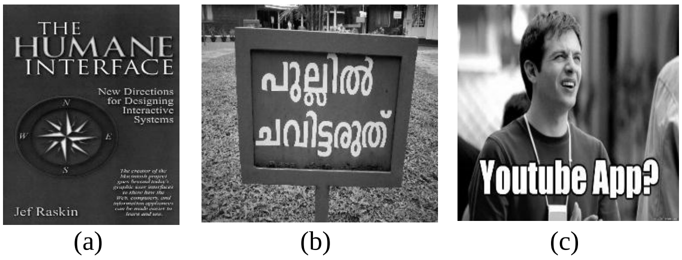
1. INTRODUCTION

Automatic identification and extraction of text regions from image is regarded as a significant and challenging research area for engineers and scientists. It is due to varying font styles, color, orientation, size and distortion of text objects with the varying lighting conditions, motion blur etc. The ultimate goal of text information extraction research, is to build systems that read any text present in various types of images with the same recognition accuracy as human but at a faster rate.

The text information within an image is used for vehicle license plate detection, analysis of articles with tables, maps and diagram and charts. It has a wide range of application in the identification of various parts in industrial automation, object identification, content based retrieval, keyword based image search, street signs, board reading, video content analysis, page segmentation, document retrieval, text based video indexing etc. [1].

Based on the nature of text information and their types, images are categorised into different groups as Document images, Scene text images and Caption text images [2]. Document images include the scanned snaps or images of any type of book covers, CD covers, documents etc. Scene text images are natural scene images with text content. The text information is overlaid or is inserted on an image to form a caption text image. The Fig.1 represents the examples of document images, scene text images and caption text images. Texts in images may show some variations with respect to the different properties such as geometry, color, motion, edge, and compression [2].

In order to understand challenges of TIE from scene images, newly considered scenes and new imaging conditions need to be detailed. New imaging conditions deal with the following parameters: raw sensor image and sensor noise, blurring effect, lighting, resolution and aliasing, outdoor/non-paper objects, scene text background, non-planar objects, unknown layout, objects in distance etc. Text has some unique characteristics depends on the spatial cohesion, orientation and frequency of the information [3].



(a) (b) (c)

Fig.1. Different types of text images. (a) Document text image (b) Scene text image (c) Caption text image

The paper consists of a quick review on various methods for localizing text from natural scene images. The paper is organized as follows: The first section gives an outline to text information extraction. The next section explains different steps involved in text information extraction. Reviews of the text information extraction methods from natural scene images are included in the third section and finally, the conclusion is enclosed.

2. STEPS INVOLVED IN TIE

Text Information Extraction (TIE) involves various steps. It includes text detection, localization, segmentation, extraction, enhancement and recognition as shown in Fig.2.

2.1 TEXT DETECTION

As there is no prior knowledge on whether or not the input image comprises any text, the presence or absence of text in the image must be determined in this phase. The text detection stage pursues the detection of the text in a given image [2]. Text detection methodologies may be divided into two main categories: (1) sliding window based approaches: Low-level features are extracted for scanning and using machine-learning techniques, each image is evaluated for the presence of text. (2) Connected components based approaches: The pixel regions with similar edge strength, color, texture or stroke width are extracted first and then text and non-text evaluation is done using rule-based or machine learning techniques [4]. Soft computing approaches are used frequently to improve the accuracy of edge detection for image segmentation [5].

2.2 TEXT LOCALIZATION

The process locates the texts within an image and bounding boxes are drawn around the texts. The text information is thus retrieved and the content information is described in an easier manner. Text localization methods are classified into two [2]: (1)

Region based method: Here, the color properties within the text region or the differences from their respective background is utilized. This approach is further divided into two: (a) Connected component (CC) based method: In this bottom-up approach, small components are combined to form the larger ones until all regions in the image are recognized. (b) Edge based method: Here, the difference between background and text is considered. For that, edges of the text boundaries are identified and are merged. Later on, numerous heuristics are used in order to filter out the non-text regions. (2) Texture based method: Here, textual properties of the text within the image that distinguish them from the background are considered. The techniques based on Gabor filters, FFT, spatial variance, Wavelet etc. are used in texture based method. The Fig.3 includes the images before and after text localization.

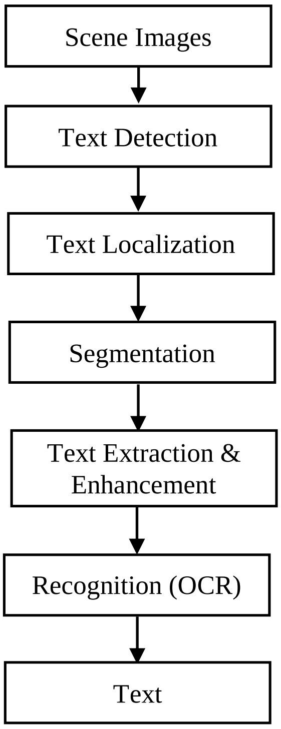


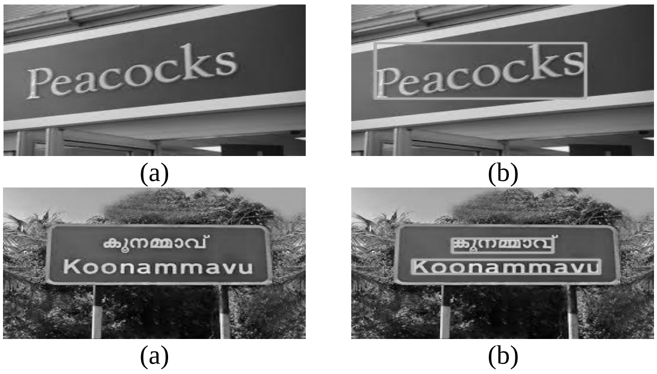
Fig.2. Steps involved in text information extraction from scene images

2.3 SEGMENTATION

The process of extracting the area of interest from a digital image is termed as segmentation. It divides an image into its constituent regions or objects. Uneven light, reflexion, shadow and low contrast are some of the challenging problems considered in this phase. [6]. Main methods includes: (1) Threshold based methods: It converts a colour or grey scale images to binary image by comparing bit intensities. It analyses the input image and finds a threshold value. The pixels above threshold are converted to white and pixels below threshold are converted to black. (2) Color cluster based methods: Focus mainly on forming clusters or group of similar pixels or segments that divides the source image into multiple similar parts [7]. (3) Statistic model based methods: Deals with the problem of complex text segmentation.

2.4 EXTRACTION

The required text is extracted from the scene images. Different techniques of extraction include Edge based text extraction, Region based text extraction, Texture based text extraction and Morphological based text extraction. In Edge based text extraction, the edges of the text boundaries are recognized and combined. Later, several methods are used to filter out the non-text regions. The properties of color in text or the change associated to background is considered in region based text extraction. Small regions are found and merged successfully as a large region. In the prior knowledge that every text has its own texture, the texture based text extraction method distinguishes the part from the background. Morphological based text extraction is a geometrical based approach for image analysis which extracts the significant text features from processed images despite of the changes in illumination conditions and text color [8].



(a) (b)

(a) (b)

Fig.3. Examples of text localization from scene images (a) Before text localization (b) After text localisation

2.5 ENHANCEMENT

Enhancement of extracted text components is inevitable due to the presence of noise and low-resolution in the image. Enhancement techniques can be categorized into two: (1) Single frame based: Single frames are used for the enhancement of images. (2) Multiple frames based: For the text extraction from the videos, as single frame is not enough for the enhancement process, the multiple frames that contains the same text content need to be considered [2].

2.6 RECOGNITION

The extracted text must be recognized in order to know what the text really means. For this, each and every character must be recognized. Later, OCR technology transforms the extracted text images into plain text. By that, the recognized characters are clearly known. The Fig.3 shows the steps in text information extraction from natural scene images.

3. TIE METHODS - A REVIEW

This review mainly focuses on various methods for localizing and extracting texts from natural scene images. The initial step of text information extraction is distinguishing the presence and absence of text in natural scene images. The proposed algorithm in [9] automatically classifies the text and non-text using Maximally Stable Extremal Region (MSER) detector, Convolutional Neural Network (CNN) method and Bag of Words (BoW) methods. K-means is applied to MSER for extracting candidate texts. Conversion of binary class to multi class classification problem is done using CNN method. The feature vector of each image is then produced by BoW along with aggregating and pooling the responses of all the candidate images. The final classification to text and non-text is made by a trained SVM classifier.

The difficulty to extract Farsi text is due to its right to left script, existence of characters with different extension in various fonts and also of different shape for the characters according to their positions within the word as in start, end, middle or standalone [13]. The hybrid approach for localizing Farsi/Arabic text is based on color based method that extracts the candidate text regions [13]. The text within the image is detected by using color and edge features. The problem of diversity in text orientation and size is solved by a new pyramid of images. The verification of candidate texts is made possible by the combination of wavelet histogram and histogram of oriented gradient. The combination of MSER Detector, Fuzzy Inference System (FIS) and Low Rank Matrix Recovery Model is an alternative method for extraction [19].

The trouble in Chinese text localization is of the intercrossed symbols having alike structures with natural scene elements [10]. Chinese text localization of scene images is based on SVM classification method [10]. The color image is processed by down-sampling, then the detection of edges is performed and later, binarization is done in the first stage. Using mathematical morphology, candidate text regions are extracted and then connected component analysis is done. Further, the PHOG-Gabor features of the candidate texts are extracted and SVM classification is employed finally. It demands higher overall performance and lower complexity. Extraction of Chinese texts from natural scenes with arbitrary orientations is also made possible with MSER Algorithm [11]. Also, MSERs pruning algorithm, self-training distance metric learning algorithm and single-link clustering algorithm together extracts the texts [12].

Recognition of Kannada text is difficult as there is no word corpus in Kannada and also has a flexional nature for the grammar [15]. Kannada text is detected, localized and is extracted from the images and digital videos [15]. After pre-processing the images based on the color reduction technique, a standard deviation based method is applied in order to detect the edges. In the final stage, the connected component properties are used for localizing the text regions. The designed system is independent of the size of the characters. Thai text from natural scene images is detected using CNN method [20]. Devanagiri characters with agglutinative and curved nature are extracted using CC Analysis [21].

Compared to regional languages, English text extraction and localization is much easier due to the nature and arrangement of the alphabets in it. It can be made possible by using boundary clustering, stroke segmentation and string fragment classification [17]. Here, Bigram color uniformity and spatial positions analyzes the text character boundaries and clustering algorithm separates them from the complex background. String fragment classification which is based on the Gabor-based text features is obtained from the feature maps of stroke distribution, stroke width and gradient. Localization of horizontal English text in complex and color images is based on proposed new projection profile [18]. G channel within the RGB color space is chosen for applying 2D db4 wavelet transform to obtain the edges of gray level image. The image is binarized and later on, the new filter is applied to remove disperses pixels and non-text area. At last, to estimate and detect the text regions, the new projection profile is applied. The alternative methods include the usage of SVM classifier, HOG Method, MSER Algorithm and the Hidden Markov Model together extracts English texts from scene images [4]. Alternative methods include Multi-scale adaptive local thresholding operator [22], Effective pruning with MSER Algorithm [23] and Stroke Width Transform (SWT) [24]. Applications of HOG and multi-scale Local Binary Pattern (msLBP) feature along with Relaxation Labeling (RL) algorithm and Markov Random Fields (MRF) model detect and localize English texts from natural scene images [25]. English text strings are detected by the combined use of Gradient and Color based methods, adjacent character and Text line grouping methods and Hough Transform [26]. Gaussian scale space pyramid concept localizes and recognizes end-to-end English scene texts [27]. English texts and numerals are detected by the combination of SVM, Conditional random field (CRF) model and Graph Cut Algorithm [28]. English text extractions are also possible with the combined use of Stroke Feature Transform (SFT), Text Covariance Descriptors for Components (TCD-C) and Text Covariance Descriptors for Text-lines (TCD-T) [29].

Multilingual language text detection and localization is done based on the texture feature extraction using first and second order statistics [14]. Firstly the texture features are extracted and later using two discriminative functions, the classification is done to obtain the candidate texts. The detected text regions are then merged and localized in the final stage. Multilingual texts are localized in the scene images by applying wavelet based edge detection and fuzzy classification [16]. Median filter is applied to the gray scale image in the pre-processing stage in order to preserve the sharp edges within the image. Discrete Haar wavelet transform characterizes the textured images. In the next stage, the Sobel operator efficiently locates the strong edges connected to the texts in the images. In the classification stage, fuzzy thresholding is applied to the images. Finally, to obtain and localize the connected components corresponding to the text regions, morphological operators are applied to the segmented image. Other methods include, CRF model for extraction [30] and SWT for localization [27]. Connected Component (CC) Analysis (Method) is combined with various other methods to extract text from natural scene images [6] [22] [24] [28] [30].

3.1 PERFORMANCE EVALUATION

The ground truth data for text localization is typically marked by bounded rectangles that comprises of gaps in the middle of characters, words and text lines. After determining the ground truth data, a decision is made on the methods to use in the matching process between localized results and ground truth data. Normally, Precision (*P*), Recall (*R*) and F_Measure (*F*) are used as the metrics for measuring the performance of algorithms in text image classification.In addition; the performance also depends on the weights assigned to false alarm or false dismissal [2]. Precision is the ratio between area of hit regions and the area of detected regions. Recall is the ratio between the area of hit regions and the area of the ground truth regions. The F_measure is obtained by combining precision and recall by harmonic mean [17]. The metrics are defined as follows:

$$\text{Precision rate } (P) = \text{TP}/(\text{TP} + \text{FP}) \quad (1)$$

$$\text{Recall rate } (R) = \text{TP}/(\text{TP} + \text{FN}) \quad (2)$$

$$\text{F\_measure } (F) = (2 * P * R)/(P + R) \quad (3)$$

Table.1. Performance Evaluation of Various Text Information Extraction Methods

Author	Year	Dataset	Methods	Work Done	P*	R*
G. Aghajari et al. [18]	2010	Own dataset of color & complex scene and video images	Wavelet based edge detection method	Localize horizontal English text appearing in images with complex backgrounds	96	91.77
Keshava Prasanna et al. [15]	2011	200 scene images & 100 MPEG video images	Color Reduction Technique, SD & CC Method	Text extraction, recognition and speech synthesis of Kannada text from digital videos and scene images with complex background	80.5	84.4
Maryam Darab and Mohammad Rahmati [13]	2012	800 scene images captured using canon camera	K-Means Cluster Algorithm, Color Based & Edge Based Methods, Wavelet Histogram, HOG & SVM Methods	Localization of Farsi text in scene images; candidate text location, verification and its classification into text and non-text class is performed	86.5	80.8
Chucai Yi and YingLi Tian [17]	2012	ICDAR 2003 & ICDAR 2011 robust reading datasets of scene images	Bigram color uniformity based method, Color Based & Edge Based Methods, Gabor based text features	Localizing English texts in scene images. Boundary clustering, Stroke segmentation, String fragment classification	81	80
Qiong XU et al. [10]	2013	400 natural scene images	Mathematical Morphology, CC Analysis, GPHOG features with SVM	Pre-processing, text coarse localization and text fine localization of Chinese text (sign boards, buildings) from scene images is done	86	83
Kumuda T and L Basavaraj [14]	2014	Own dataset along with ICDAR 2011 dataset	First and second order statistical properties	Texture feature extraction, candidate area detection, merging and localization from scene images of multilingual languages with different font size, color and orientation	90.27	86.91
Chengquan Zhang et al. [9]	2015	7302 text images & 8000 non-text images from internet	MSER Detector, CNN method, BoW Method	Scene image classification to text and non-text; images with different languages, colors, fonts, scales, orientation and layout are used	91.6	90.8
Shivananda V. Seeri et al. [16]	2015	238 scene images (ICDAR dataset included)	Haar wavelet, Sobel edge detector, K-Means Cluster Algorithm, Mathematical Morphology, Fuzzy classification	Multilingual text localization and non-text removal from scene images	79.54	89.21

* Evaluation measures (P and R) given are based on the research papers reviewed

where, True Positive (*TP*) is the number of correctly classified text images, False Positive (*FP*) is the number of incorrectly classified non-text images and False Negative (*FN*) is the number of incorrectly classified text images [9]. We reviewed more than 50 research papers of different languages. Among that, 25 papers are chosen for detailed study. Out of that, Table.1 summarizes those which reported a high precision and recall rate above 80%. Table also includes various methods, features, dataset, year of publication, and the precision and recall rates evaluation.

4. CONCLUSION

For content-based image analysis, the extraction of text information from natural scene images is inevitable. Achieving systems that can read any type of texts present in numerous kinds of images with the same recognition accuracy as humanoid but at a quicker rate is the ultimate goal of text information extraction research. The aspects backing to the intricacy includes complex background, uneven lighting, variations in texts, font, size and text-line orientation. It is noticed that not many research works have been carried out specifically for the extraction of Indian language texts due to its agglutinative and curved nature. In comparison, English characters have a much better and easier style of representation. As it is detected in an easier manner, several research works are done in this area. More studies are required to improve the recognition rates in order to deal with the natural scene images. Also, more research works are desired to be done in regional languages.

REFERENCES

[1] Shilpi Rani and Kanchan, “An Overview of Text Extraction from Colored Images”, *MIT International Journal of Computer Science and Information Technology*, Vol. 5, No. 1, pp. 1-4, 2015.

[2] Keechul Jung, Kwang In Kim and Anil K. Jain, “Text Information Extraction in Images and Video: A Survey”, *Pattern Recognition*, Vol. 37, No. 5, pp. 977-997, 2004.

[3] Victor Wu, Raghavan Manmatha, and Edward M. Riseman, “Text Finder: An Automatic System to Detect and Recognize Text in Images”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 21, No. 12, pp. 1-35, 1999.

[4] Arpit Jainz, Xujun Peng, Xiaodan Zhuang, Pradeep Natarajan and Huaigu Cao, “Text Detection and Recognition in Natural Scenes and Consumer Videos”, *Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing*, pp. 1254-1258, 2014.

[5] N. Senthilkumaran and R. Rajesh, “Edge Detection Techniques for Image Segmentation - A Survey of Soft

Computing Approaches", *International Journal of Recent Trends in Engineering*, Vol. 1, No. 2, pp. 250-254, 2009.

[6] Xiaopei Liu, Zhaoyang Lu, Jing Li and Wei Jiang, "Detection and Segmentation Text from Natural Scene Images Based on Graph Model", *WSEAS Transactions on Signal Processing*, Vol. 10, No. 1, pp. 124-135, 2014.

[7] Monika Xess and S. Akila Agnes, "Survey on Clustering Based Color Image Segmentation and Novel Approaches to FCM Algorithm", *International Journal of Research in Engineering and Technology*, Vol. 2, No. 12, pp. 346-349, 2013.

[8] A.J. Jadhav, Vaibhav Kolhe and Sagar Peshwe, "Text Extraction from Images: A Survey", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 3, No. 3, pp. 333-337, 2013.

[9] Chengquan Zhang, Cong Yao, Baoguang Shi and Xiang Bai, "Automatic Discrimination of Text and Non-Text Natural Images", *Proceedings of 13th International Conference on Document Analysis and Recognition*, pp. 886-890, 2015.

[10] Qiong Xu, Zongliang Gan, Changhong Chen and Feng Liu, "Novel Chinese Text Localization Method for Natural Images through SVM Classification", *Journal of Computational Information Systems*, Vol. 9, No. 18, pp. 7291-7298, 2013.

[11] Lluis Gomez and Dimosthenis Karatzas, "Multi-script Text Extraction from Natural Scenes", *Proceedings of 12th International Conference on Document Analysis and Recognition*, pp. 1-5, 2013.

[12] Xu Cheng Yin, Xuwang Yin, Kaizhu Huang and Hong Wei Hao, "Robust Text Detection in Natural Scene Images", *IEEE Transactions Pattern Analysis and Machine Intelligence*, Vol. 36, No. 5, pp. 970-983, 2013.

[13] Maryam Darab and Mohammad Rahmati, "A Hybrid Approach to Localize Farsi Text in Natural Scene Images", *Proceedings of International Neural Network Society Winter Conference*, Vol. 13, pp. 171-184, 2012.

[14] T. Kumuda and L Basavaraj, "A Novel Technique for Text Detection and Localization in Natural Scene Images", *International Journal of Engineering Research and Technology*, Vol. 3, No. 4, pp. 2675-2680, 2014.

[15] Keshava Prasanna, Ramakhanth Kumar P, Thungamani.M and Manohar Koli, "Kannada Text Extraction from Images and Videos for Vision Impaired Persons", *International Journal of Advances in Engineering and Technology*, Vol. 1, No. 5, pp. 189-196, 2011.

[16] Shivananda V. Seeri, J.D. Pujari and P.S. Hiremath, "Multilingual Text Localization in Natural Scene Images using Wavelet based Edge Features and Fuzzy Classification", *International Journal of Emerging Trends and Technology in Computer Science*, Vol. 4, No. 1, pp. 210-218, 2015.

[17] Chucai Yi and Yingli Tian, "Localizing Text in Scene Images by Boundary Clustering, Stroke Segmentation, and String Fragment Classification", *IEEE Transactions on Image Processing*, Vol. 21, No. 9, pp. 4256-4268, 2012.

[18] G. Aghajari, J. Shanbehzadeh and A. Sarrafzadeh, "A Text Localization Algorithm in Color Image via New Projection Profile", *Proceedings of the International Multi Conference of Engineers and Computer Scientists*, Vol. 2, pp. 1-4, 2010.

[19] Shaho Ghanei and Karim Faez, "Localizing Scene Texts by Fuzzy Inference Systems and Low Rank Matrix Recovery Model", *Computer Vision and Image Understanding*, Vol. 142, pp. 94-110, 2016.

[20] Thananop Kobchaisawat and Thanarat H. Chalidabhongse, "Thai Text Localization in Natural Scene Images using Convolutional Neural Network", *Proceedings of Annual Summit and Conference on Asia Pacific Signal and Information Processing Association*, pp. 1-7, 2014.

[21] Vishwanatha Kaushik and C.V. Jawahar, "Detection of Devanagari Text in Digital Images using Connected Component Analysis", *Proceedings of National Conference on Document Analysis and Recognition*, pp. 41-48, 2003.

[22] Xiaoqian Liu, Ke Lu and Weiqiang Wang, "Effectively localize Text in Natural Scene Images", *Proceedings of 21st International Conference on Pattern Recognition*, pp. 1197-1200, 2012.

[23] Lukas Neumann and Jiri Matas, "Text Localization in Real-World Images using Efficiently Pruned Exhaustive Search", *Proceedings of International Conference on Document Analysis and Recognition*, pp. 687-691, 2011.

[24] Boris Epshtein, Eyal Ofek and Yonatan Wexler, "Detecting Text in Natural Scenes with Stroke Width Transform", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2963-2970, 2010.

[25] Yi Feng Pan, Xinwen Hou and Cheng Lin Liu, "A Robust System to Detect and Localize Texts in Natural Scene Images", *Proceedings of 8th International Workshop on Document Analysis Systems*, pp. 35-42, 2008.

[26] Chucai Yi and Ying Li Tian, "Text String Detection from Natural Scenes by Structure-based Partition and Grouping", *IEEE Transactions on Image Processing*, Vol. 20, No. 9, pp. 2594-2605, 2011.

[27] Lukas Neumann and Jiri Matas, "On Combining Multiple Segmentations in Scene Text Recognition", *Proceedings of 12th International Conference on Document Analysis and Recognition*, pp. 523-527, 2013.

[28] Li Rong, Wang Suyu and Zhixin Shi, "A Two Level Algorithm for Text Detection in Natural Scene Images", *Proceedings of 11th International Workshop on Document Analysis Systems*, pp. 329-333, 2014.

[29] Cong Yao, Xiang Bai, Wenyu Liu, Yi Ma and Zhuowen Tu, "Detecting Texts of Arbitrary Orientations in Natural Images", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1083-1090, 2012.

[30] Yi-Feng Pan, Xinwen Hou, and Cheng-Lin Liu, "A Hybrid Approach to Detect and Localize Texts in Natural Scene Images", *IEEE Transactions on Image Processing*, Vol. 20, No. 3, pp. 800-813, 2011.