Pertanika J. Sci. & Technol. 27 (4): 1759 - 1781 (2019)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Detection and Recognition via Adaptive Binarization and Fuzzy Clustering

Saad Mohmad Saad Ismail, Siti Norul Huda Sheikh Abdullah and Fariza Fauzi*

Center of Cyber Security, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM, Bangi, Selangor, Malaysia

ABSTRACT

Detection and identification of text in natural scene images pose major challenges: image quality varies as scenes are taken under different conditions (lighting, angle and resolution) and the contained text entities can be in any form (size, style and orientation). In this paper, a robust approach is proposed to localize, extract and recognize scene texts of different sizes, fonts and orientations from images of varying quality. The proposed method consists of the following steps: preprocessing and enhancement of input image using the National Television System Committee (NTSC) color mapping and the contrast enhancement via mean histogram stretching; candidate text regions detection using hybrid adaptive segmentation and fuzzy c-means clustering techniques; a two-stage text extraction from the candidate text regions to filter out false text regions include local character filtering according to a rule-based approach using shape and statistical features and text region filtering via stroke width transform (SWT); and finally, text recognition using Tesseract OCR engine. The proposed method was evaluated using two benchmark datasets: ICDAR2013 and KAIST image datasets. The proposed method effectively dealt with complex scene images containing texts of various font sizes, colors, and orientation; and outperformed state-of-the-art methods, achieving >80% in both precision and recall measures.

Keywords: Adaptive binarization; fuzzy C-means; image enhancement; statistical and geometrical features; text detection; text extraction

ARTICLE INFO

Article history:
Received: 30 January 2019
Accepted: 03 July 2019
Published: 21 October 2019

E-mail addresses:
ababnah2010@gmail.com (Saad Mohmad Saad Ismail)
snhsabdullah@ukm.edu.my (Siti Norul Huda Sheikh Abdullah)
fariza.fauzi@ukm.edu.my (Fariza Fauzi)
* Corresponding author

INTRODUCTION

With the growing popularity of mobile/ wearable devices with embedded digital cameras, an abundance of scene images is available. It is common to find text present/

ISSN: 0128-7680
e-ISSN: 2231-8526

embedded in scene images (scene text) such as shop signage, street names and billboards which can contribute significant information about a particular scene. Text has been and still is the most intuitive way for most people to represent information. As such, extracting text from images of real scenes has become an important task especially with the advent of life changing technologies such as the Internet of Things, robotics, automation and augmented reality. Scene text detection and extraction play an important role in many applications such as detecting texts on signposts (Raghunandan et al., 2018), billboards, nameplates, labels, license plate recognition (Huang et al., 2018; Li et al., 2018), invoice number detections (Sun et al., 2019) , mobile text reader, ticket number reader (Kraisin & Kaothanthong, 2018), handphone number reader, scene-to-text conversion and pronunciation for visually impaired people (Qaisar et al., 2019).

Natural scene images are usually taken using digital and mobile cameras. These photos are diverse, of varying quality (Nagaraju et al., 2015; Risnumawan et al., 2014) making the detection and extraction of scene text a difficult and challenging task. Various issues include changes in light intensity, text alignment, font size, color, camera angles, occlusion and so on (Behzadi & Safabakhsh, 2018; Ning et al., 2015; Raghunandan et al., 2018; Wang et al., 2015; Zhang et al., 2013). Generally, text extraction and recognition techniques can automatically detect areas of texts in a scene image by marking boundaries (usually bounding boxes) and reading its text content as a series of (Unicode) characters. The digitized text can then be further processed by a computer. Extracting text from real scene images generally consists of three phases: (1) detection and localization, (2) enhancement and segmentation, and (3) optical character recognition (OCR). The most difficult problem when dealing with natural text extraction and recognition is due to the differences in the font, color, alignment of text size, change in lighting, and reflections (Jadhav et al.,2013; Sulaiman et al., 2019). These persistent challenges have made histogram-based segmentation methods (Behzadi & Safabakhsh, 2018; Ning et al., 2015; Raghunandan et al., 2018; Wang et al., 2015; Zhang et al., 2013) and handcrafted approaches to be trivial and insignificant in such scenario or situation.

Many researchers have proposed several text extraction methods from complex images (Sumathi et al., 2012). Although each method has its own unique approach and objectives, they have one common goal which is to obtain the best extraction of text from an image in order to maximize text recognition. An in-depth literature review on OCR has been presented in (Jung et al., 2004; Zhang et al., 2013) whereby each review highlighted its own research gaps, benchmark dataset, evaluation criteria, drawbacks and future direction. Jung et al. (2004) gave a comprehensive survey in classifying algorithms proposed to address the related problems in text extraction. A typical text extraction process involves detection, localization, tracking, extraction, enhancement, and recognition of the text from any scene image. Karanje and Dagade (2014) focused on the advantages and disadvantages of the major categories of text extraction approaches. They also described state-of-the-art

methods for the detection of text, text segmentation, and character recognition in natural scene images. Samadhiya and Khatri (2014) reviewed various text extraction algorithms and discussed the performance evaluation and challenges. The authors offered a collection of recent techniques for text information extraction.

Besides that, several researches implemented the K-means clustering algorithm to perform segmentation and recognition. Burney and Tariq (2014) proposed K-means clustering for segmentation where dissimilar pixels from the source image were divided into several (or a predetermined number) similar regions or segments. K-means clustering can also be used directly for classification. It distinguishes the objects regions into text or non-text (Wang et al., 2011). On the other hand, the drawback of K-means clustering approach in text extraction and segmentation is that it does not perform well when dealing with natural image pixels that have very low or very high contrast. This weakness causes the separation of text from its background to become nontrivial.

Shivakumara et al. (2011) described a Laplacian operator based on the method of frequency domain model. In this approach, the input image was filtered using Fourier-Laplacian transform. Then it used K-means clustering to identify regions of candidate text on the basis of the maximum difference. To overcome the limitation of K-means, a new iterative nearest neighbor symmetry has been introduced intentionally for restoring missing text whereas fixing window based on angular relationship relying on sub-bands and its fused bands is proposed to improve arbitrary oriented text in the natural scene environment (Raghunandan et al. 2018). Apart from classic natural scene-handcrafted methods, several researchers were also discovering machine-crafted natural text segmentation by introducing word-fence and fully convolutional Densenet (Behzadi & Safabakhsh, 2018) whilst Huang et al. (2018) presented a unified end-to-end trainable deep network, which could simultaneously locate and recognize vessel plate number.

Phan et al. (2009) developed an approach to detect text with the Laplacian operator. Then K-means is used to classify all pixels in clusters. Based on the literature, less attention was given to other clustering approaches, such as fuzzy c-means (FCM) clustering. FCM is a frequently used clustering method in pattern recognition which allows one piece of data to belong to two or more clusters (Horvath, 2006; Rajaby et al., 2016). As such, it is investigated in this paper for the tasks of text detection and extraction.

The remainder of the paper is organized as follows: The Method section describes the overall proposed method consisting of the preprocessing and enhancement phase, followed by the hybrid adaptive segmentation and clustering phase to detect, extract and recognize text. The Results and Discussion section examines measurements of extraction and recognition of different datasets that are used to objectively compare the proposed method to existing text extraction methods. Experimentally, these different methods are compared using different datasets with different environment types. Finally, a conclusion is drawn.

METHOD

The block diagram of the proposed method is illustrated in Figure 1. Given an input image, text color image remapping was performed using NTSC color distance and histogram normalization to obtain a clean grayscale image with clear and high contrasting background and text, addressing the poor/low contrast issue for the effective recognition. Then, to convert the clean grayscale image into a binary image, a hybrid approach of clustering and adaptive binarization was performed. The clustering of grayscale image into two clusters was achieved using FCM. The output was combined with the outputs of the adaptive binarization of color image channels and the complement of the binary image. The aim of this step is to ensure the conversion is performed without fading any text regions with high appearance, in addition to unifying the background for all text regions within the same image. Finally, the regions were extracted using a connected component technique based on several statistical distributions and shape features. The extracted text regions were filtered using a rule-based technique and stroke width transform using Manhattan distance transform. For character recognition, the open-source Tesseract OCR engine was employed. The subsequent sections will detail out each stage in the proposed method.

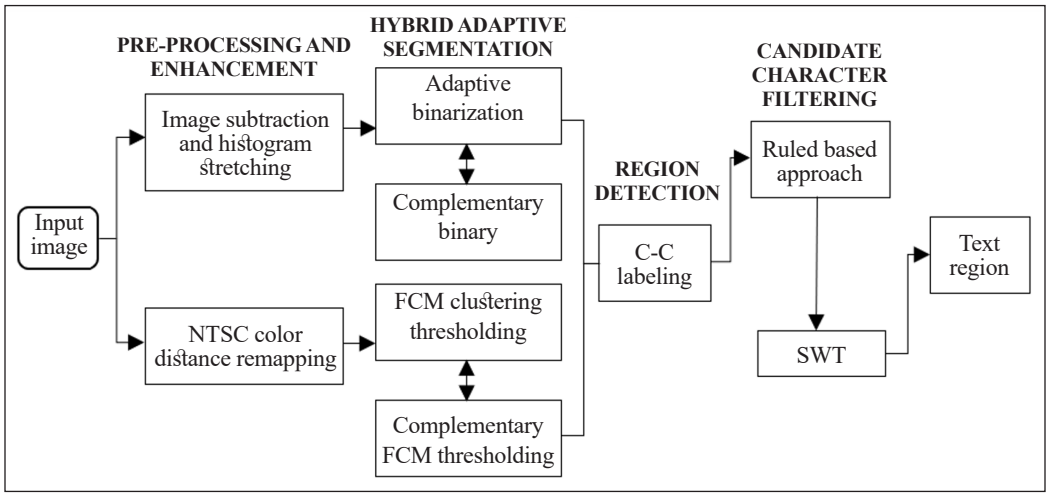


Figure 1. Block diagram of the proposed approach

Preprocessing and Enhancement

Natural scene images can contain noise for example, salt and pepper noise, impulse noise etc. or they can be blurred due to inadvertent movement of camera, poor lighting condition causing over- or under-exposure, light reflection and non-uniform color. For that purpose, image preprocessing and enhancement techniques are applied to reduce such noise. In the proposed framework, color distance remapping and low contrast enhancement by subtraction and normalization using histogram stretching are performed.

Color Space Mapping. The proposed image color enhancement method is aimed to suppress the reflective light, uneven illumination in noisy scene images. Hence, the source color image is enhanced by using the proposed exploited NTSC color mapping method. In general, there are different image color spaces such as RGB, NTSC, HSV and each one has its own unique features. Here, the NTSC color space was selected primarily for its ability to characterize (Shah & Thakar, 2010) and separate gray color from other colors (red and green). Ultimately, this exploited remapping process can easily tackle the aforementioned problems.

In the NTSC format, image data consists of three components: luminance (Y), and chrominance: hue (I), and saturation (Q). The first component, luminance, represents the grayscale information, while the last two components represent chrominance (color information). Usually, background pixel values are mostly classified in luminance component. By multiplying these values with a small value, $\propto$, the gray (luminance) values decreases to merely zero or black. Contrariwise, multiplying the chrominance by big values (β,γ) the appearance of foreground (object) increases to merely white. Heuristically, the best $\propto$,β,γ values are obtained. Equation 1 denotes the NTSC color distance remapping function where luminance and chrominance are multiplied by $\propto$,β,γ.

$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.299 & 0.596 & 0.211 \\ 0.587 & -0.274 & -0.523 \\ 0.114 & -0.322 & 0.312 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \times \begin{bmatrix} \propto \\ \beta \\ \gamma \end{bmatrix}, \quad (1)$$

where: $\propto$ =0.6, β=8, γ=16.

This mapping process solves the light reflection and low contrast flaws in scene image. One main benefit of this format is that the grayscale information is separated from color data, therefore, both color as well as black and white sets can share the same channel.

Figure 2 illustrates the result of the proposed NTSC color mapping method: the original image, shown in Figure 2(a) is compared to images in Figure 2(b)-2(d) which are processed using regular grayscale conversation / RGB channels and images in Figure 2(e)-2(g) are processed using the proposed NTSC color distance remapping function. As demonstrated, the proposed method effectively differentiated the text from background, resulting in a cleaner image obtained from the scene image in Figure 2(e)-2(g), with the exceptions of scene images having a very low contrast with a high light reflection at the same time, or when there are highly-effective shadows in the text region. Ultimately, the proposed NTSC enhancement step addresses the low contrast and illumination problems, providing cleaner images to the subsequent step (Hybrid Adaptive Segmentation) of the overall method which aims to efficiently detect and extract text from scene images.

Figure 2. An example showing the effectiveness of the proposed image enhancement. method: (a) original image (b-d) before and (e-g) after introduction of the proposed exploited NTSC for image enhancement method according to RGB channels (red, green, blue, respectively).

Hybrid Adaptive Segmentation and Clustering

The proposed hybrid adaptive segmentation method combines the results of two segmentation techniques: FCM clustering-based segmentation and adaptive binarization-based segmentation. The hybrid approach draws on the strengths of FCM clustering in segmentation and adaptive binarization in addressing low contrast and illumination problems, which is essentially a drawback in FCM clustering. This results in an improved segmentation method which retains most of the text regions with minimal noise.

Segmentation using Proposed Fuzzy C-Means (FCM) Clustering. For text extraction from scene images, the FCM clustering was utilized for binarization purposes. First, the grayscale image was clustered into two categories, and then a global threshold method was applied to set pixels in each cluster to either black or white. The binarization process utilizing FCM clustering is described in Algorithm 1 which was used in Horvath (2006).

The algorithm assumes the scene image has been converted into the grayscale feature space with C number of clusters and the stop condition and fuzziness parameter are denoted by ε and M respectively. Here, it is used for image binarization, hence C was set to 2 to segment out the text from the background.

Algorithm 1. The Proposed FCM Clustering Method for Image Binarization

Input: Grey Image
Output: Binary Image
Start
Step 1: Cluster image in the grayscale feature space, with the following conditions: number of clusters = C, fuzziness index = f, stop condition = ε.
Step 2: Repeat for each pixel $I(i,j)$ of image I.
Step 3: Compute which cluster C does pixel $I(i,j)$ belong to.
Step 4: Identify if there exist a segment R_k whose points belong to the same cluster C, in the closest surroundings of pixel $I(i,j)$.
Step 5: If segment R_k exists, then pixel $I(i,j)$ is added to segment R_k, else create a new segment R_n and add pixel $I(i,j)$ to the new segment R_n.
Step 6: Merge all segments which belong to one cluster and are neighbors.
Step 7: Arrange borders of all segments.
Step 8: The pixels belonging to cluster one are set to white and the pixels belonging to cluster two are set to black
End

The experiment results indicated that the FCM clustering technique gave good results for many cases. However, the drawback of this method is that it is affected by strong lighting and non-uniform text color that cannot be overcome. Figure 3 presents an example of the binary segmentation results using FCM clustering algorithm.

Segmentation using Proposed Adaptive Binarization. Adaptive compound binarization is the second segmentation technique proposed to segment and binarize text from scene images. The adaptive binarization is expressed in the equation below:

$$T_{Adaptive} = \mu_l * (1 - T_{G\ Dynamic}) \qquad (2)$$

where $T_{Adaptive}$ is the local threshold value and μ_l is the local mean value of integral image pixels estimated based on a dynamic generation window size ($T_{G\ Dynamic}$). Based on this $T_{Adaptive}$ value, the binarization process is defined in equation (3). With the assumption that $I_{binary}\ (x, y)$ is the binary image and $I(x, y)$ is the grayscale image, therefore when the current pixel $I(x, y)$ is less than $T_{Adaptive}$, it is set to black, otherwise it is set to white.

$$T_{binary}(x,y) = \begin{cases} \text{black}, I(x,y) & < T_{Adaptive} \\ white & otherwise \end{cases} \quad (3)$$

As mentioned in the beginning of the section, the proposed adaptive binarization compensates the drawbacks of FCM clustering method wherein its poor ability to segment images for very low contrast or high illuminance e.g. texts “A 12” and “A 120” appear as a black box, “PR” is missing, “I” is broken and missing text “inhabited environment” in Figure 3(b). Figure 3(c) presents the results of the proposed adaptive compound binarization method on the original set of images in Figure 3(a). It is demonstrated in Figure 3(c) that adaptive binarization produces better results in extracting text from low contrast or high illuminance scene images. However, it is also observed that many unnecessary background details (noise) are also present. Hence, a hybrid adaptive segmentation step is recommended in the overall proposed approach for text detection and recognition from degraded scene images.

Figure 3. Sample segmentation results. (a) Original image set (b) segmentation results using FCM clustering and (c) segmentation results using adaptive compound binarization.

Hybrid Segmentation. Final segmentation was the aggregated results of FCM clustering for the grayscale image and its compliment with the results of the proposed adaptive binarization of the grayscale image and its compliment. This process gives the best segmentation that is capable of preserving text region in scene images with lowest noise. The following equation denotes the hybrid segmentation process.

$$I_{hybrid} = I_{binary} + I^{c}_{binary} + U + U^{c} \qquad (4)$$

where I_{binary} is the adaptive binarization image c.f. equation (3), I^{c}_{binary} is the complement adaptive binarization image, U is the FCM binary image and U^{c} the complement FCM binary image. Figure 4 shows some example results of the proposed hybrid segmentation step.

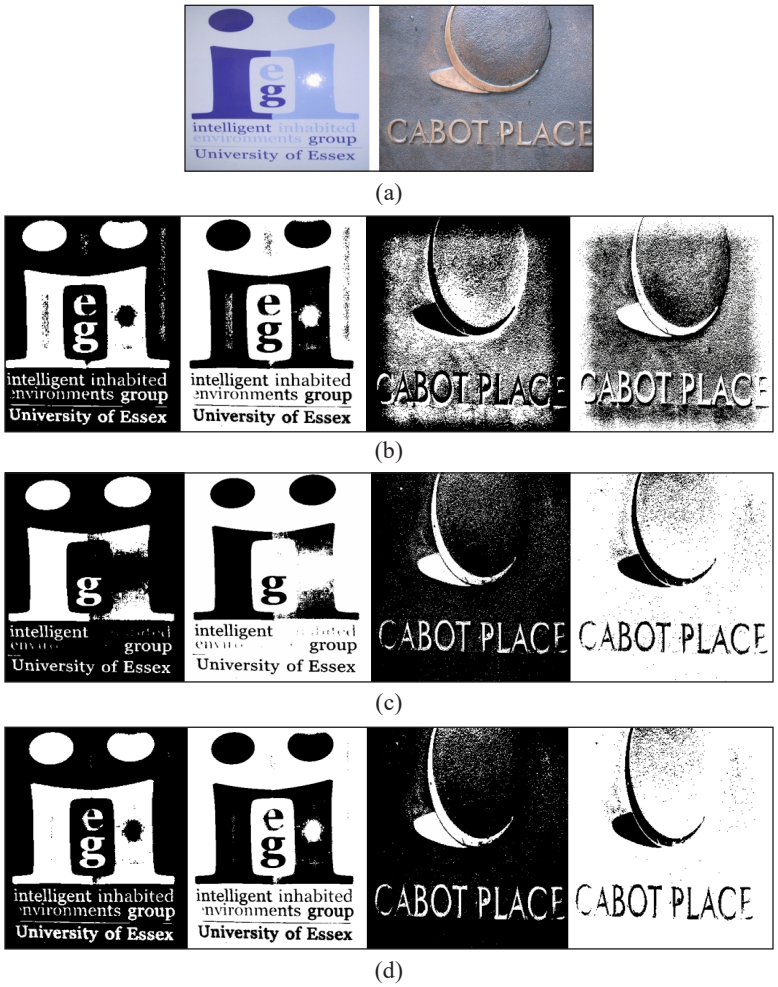


Figure 4. Example results of the proposed hybrid segmentation method: (a) original images (b) adaptive compound binarization and its complement (c) FCM segmentation and its complement (d) Final segmentation for the original images and their complement.

Text Extraction from Binary Image

Preserving all the text information together with other objects as the foreground is essential. After applying the proposed hybrid segmentation step, a binary image was produced. All the foreground objects were then, labelled using connected component method (Nosal, 2008). The challenge now was to remove the non-text objects from the foreground in the text extraction phase. The text extraction phase was divided into two stages: local character and text region filtering. In the first stage, a rule-based local character filtering was implemented using shape features whereas text region filtering used Stroke Width Transforms (SWT) to eliminate non-text objects.

CC-based methods typically use a bottom-up approach by grouping small components into successively larger components until all regions are identified in the image (Kumar et al., 2010; Raj & Ghosh, 2014; Yao et al., 2007). Statistical and geometrical analyses are required to merge the text components based on the spatial arrangement of those components. In the proposed method, seven geometric and shape features, tabulated in Table 1, are considered in the overall proposed detection and extraction method for localizing the text regions.

Table 1
List of features used in the rule-based local character filtering for text localization

Feature	Definition
Eccentricity of the region	Eccentricity = $\sqrt[2]{1-(\frac{a}{b})^2}$
Relative Smoothness	$1-\frac{1}{1+\sigma^2}$, $\boldsymbol{\sigma}$ is the standard deviation
Area	$Area\ ratio = \frac{Region\ pixels}{Total\ image\ pixels}$
Perimeter	Perimeter = $\sum_{i=0}^{m-1}$ Length c_i, $c_i = (0 \ldots m-1)$
Length and Width of the Region	Area high and width size
Aspect ratio	The ratio of pixels in the region calculated by dividing the width by the height of the region.
Orientation	Angle Between the x Axis and Major Axis of the Region

Proposed rule-based Approach for Local Character Filtering. For text detection and extraction, a set of hierarchical rules was developed, in relation to the above-mentioned geometric features to filter out the non-text areas from the CC regions or CC listing as described in Algorithm 1 The degree of variability for each feature is shown in Table 2 and Table 3.

Table 2
Degree of geometrical and shape features used for constructing the local character filtering rules

No.	Feature	Low	Normal	High
1	Eccentricity	<0.2	0.2 to 0.6	>0.8
2	Area	<70	70 to 2000	>10000
3	Relative Smoothness	<0.01	0.01 to 0.06	>0.08
4	Perimeter	<20	20 to 100	>200
5	Length and Width	<50	50 to 200	>200
6	Aspect ratio	<1	1 to 5	>7
7	Orientation	0^0 to 35^0	-90^0 to 35^0	70^0 to 90^0

*Please take note that for each row vertical represents conjunction (AND operator) and horizontal represents conjunction (OR operator) and between rows represents (OR operator)

Table 3
Decision table for non-character region

No.	Feature	Condition		Decision
		Low	High	
1	Eccentricity		1	Rule 1
2	Area	1	1	
3	Relative Smoothness		2	Rule 2
4	Orientation	2		
5	Length and Width	3	3	Rule 3
6	Aspect ratio		4	Rule 4
7	Perimeter	4	4	

*Please take note that for each row vertical represents conjunction (AND operator) and horizontal represents conjunction (OR operator) and between rows represents (OR operator)

Algorithm 2: The proposed rule-based approach for text regions filtering

Input: Child binary image of labelled component
Output: character or non-character regions
Begin
Step 1: if (Eccentricity = high **and** Area = low **or** high) **or** *//Rule 1*
Step 2: (Relative Smoothness = high **and** the Orientation = low) **or** *//Rule 2*
Step 3: **(**Length and Width = high **or** low) **or** *//Rule 3*
Step 4: **(**Aspect ratio = high **and** Perimeter = high **or** low) *//Rule 4*
Step 5: **then** Region = non-character
Step 6: **else** Region = character.
End

The proposed rules give an efficient tool for filtering a text object from non-text. They can extract all types of texts from scene images, regardless of the differences between the texts in terms of size, font type, orientation, and the proximity to other objects.

Text Region Filtering using SWT. In this stage, SWT method proposed by Chen, et al. (2011) was employed. The SWT algorithm is based on the idea that stroke is central in text, be it handwritten or typed. Stroke width typically refers to the length of a perpendicular line that connects a text edge pixel to another pixel on the opposite text edge (distance between two pixels on the two parallel text edges). The stroke width is almost consistent throughout a single character in contrast to not-text regions where there is significant change in the stroke width due to their irregularities. In SWT, the first step is to obtain the skeletons of the binary image. For each foreground pixel in the skeleton, distance transform is applied to compute the Manhatten distance from the pixel to the nearest its boundary/edge. This results in a skeleton-distance map. The standard deviation (STD) is then calculated on the skeleton-distance map of each CC strokes to compute the difference between the true text regions and false positives. It should be noted that text characters typically have a much smaller STD compared to the false positive which is fixed. Based on this property, CCs with large STD are removed.

The algorithm of SWT method using Manhattan distance transform for text filtering is expressed in the following equation:

$$M_{distance} = |(x_1 - x_2)^2 + |(y_1 - y_2)^2| \qquad (5)$$

Algorithm 3: The Proposed Stroke Width Transform for Character Filtering

Input: Child Binary Image of a group connected components
Output: Character Region
Start

Step 1: Calculate Manhattan distance transform ($M_{distance}$)

Step 2: Calculate stroke width using ($M_{distance}$).

Step 3: For each proposed connected component label (CC), calculate the local standard deviation (σ_{stroke}) for each region strokes.

Step 4: Filter the Character by the following threshold equation:

$$T_{extract}(x,y) = \begin{cases} \text{black}, \sigma_{stroke}) > 1 \\ \textit{white, otherwise} \end{cases} \qquad (6)$$

End

Figure 5 illustrates an example result of the proposed SWT method which is used to filter all non-character candidates that were not eliminated in the first filtering step.

RESULTS AND DISCUSSION

The proposed method is evaluated in two ways: visual and analytical. The visual evaluation aims to demonstrate the effectiveness of the proposed method by presenting the visual results for various types of text detection and extraction challenges whilst the analytical evaluation is based on one or more of the benchmark metrics adopted since DIBCO 2009 [Precision and recall, Picture Signal-to-Noise Ratio (PSNR) and Negative Rate Metric (NRM)] to measure the robustness of the proposed method. Two benchmark datasets were used in the evaluations. Firstly, the visual test evaluated the text extraction results by comparing against different state-of-the-art methods. Secondly, the analytical test evaluated the performance of the proposed framework in terms of recognizing the characters and texts (Optical Character Recognition, OCR).

Visual Experiment

The ICDAR Reading Competition organized by the International Conference on Document Analysis and Recognition (ICDAR), has been held five times, in 2003, 2005, 2011, 2013 and 2015. The dataset used in this text reading competition contains about 233 images with ground truth. The KAIST scene text dataset consists of 3000 images captured in various environments, which includes both indoor and outdoor scenes under varying lighting conditions (clear day, night, strong artificial lights). The KAIST dataset is used for evaluation purposes to ensure the coverage of scene text images with different environments and lighting conditions. In addition to the ICDAR and KAIST datasets used for evaluation in this study, we randomly chose several images from another dataset of poor quality such as a degraded document image and a noisy image of license plate. These images are used in the visual assessment for demonstrating the effectiveness of the proposed method in extracting text. Figures 5 to 7 highlights some examples of text extraction results from images that are very difficult to extract.

Figure 8 presents a sample results of the recognition process. A visual analysis of the results demonstrates the effectiveness of proposed OCR framework in accurately recognizing the text particularly from some difficult scene text images containing multi-oriented texts, which the other state-of-the-art methods such as Neumann and Matas (2015) and Gomez and Karatzas (2013) were not able to recognize. The improved results are due to the fact that the proposed text detection and extraction method includes a pre-processing step to prepare and correct difficult images, enhancing it for recognition.

Analytical Experiment

To evaluate the proposed extraction and recognition methods, a comparative study with different state-of-the-art methods in text extraction and recognition was conducted, including the current top ranked methods in ICDAR competition such as Bai et al. (2013),

Gomez and Karatzas (2013), Kumar and Lee (2010), Text Spotter by Neumann and Matas (2015), Shi et al. (2013), Sung et al. (2015), Novikova et al. (2012) and Yin et al. (2014). The proposed extraction and recognition methods were evaluated separately. The precision and recall metrics were computed to evaluate the performance of the extraction method whereas accuracy metric was used to evaluate the recognition method. Evaluation was carried out using two benchmark datasets: ICDAR Robust Reading 2013 dataset and KAIST scene image dataset.

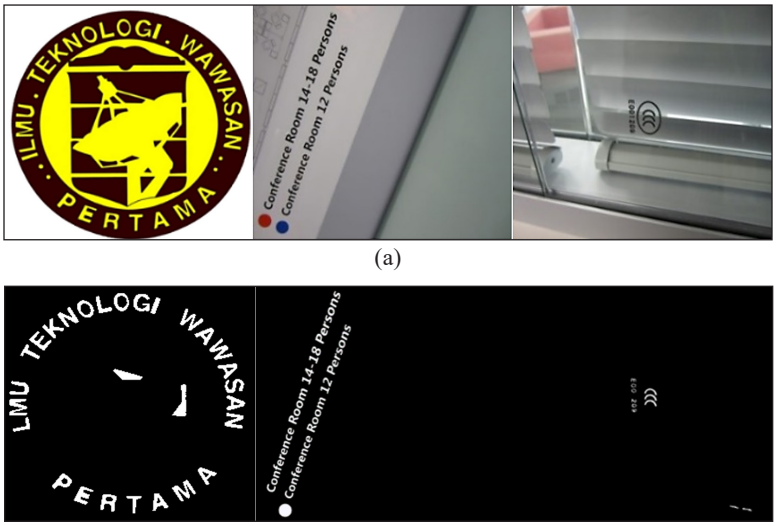
Precision and recall evaluation measures are commonly applied in the area of text extraction (Yin et al., 2014; Neumann & Matas, 2015). Precision is described as the number of text correctly extracted divided by the total number of extracted text and recall is described as the number of text correctly estimated divided by the total number of ground truths (the words found in the original scene image). The precision and recall metrics for text extraction are computed as follows:

$$Precision = \frac{N_{tp}}{N_{tp} + N_{fp}}, \; Recall = \frac{N_{tp}}{N_{tp} + N_{fn}} \qquad (7)$$

where N_{tp}, N_{fp}, and N_{fn}, denote the number of True Positive, False Positive, and False Negative values, respectively.

(a) (b)

Figure 5. Examples of results of the proposed method for images classified as very complex for extraction: (a) original image set on the left side and (b) text extraction results on the right. (*Red and green line border indicates non-text and text regions subsequently)



(a)

(b)

Figure 6. Some example results for non-horizontal text scene image: (a) Original image set and (b) text extraction results

(a)

(b)

Figure 7. Examples of challenges in ICDAR 2013 images and its results using the proposed text detection and extraction methods: (a) original image set and (b) extraction results

(a) (b)

Figure 8. Sample output of the recognition process using the proposed method for more complex images: leftmost image is the original image; center image is the extracted text and rightmost image is the recognition result. (a) Slanted text art shape with complex background, and (b) art text shape from KAIST/ICDAR datasets and google image search engine.

The aim of performance assessment for a text detection and extraction method is to measure the difference between the expected text result and the real text result of the method. Table 4 shows the results for the proposed text detection and extraction methods, comparing it with Sung et al. (2015), Yin et al. (2014), Bai et al. (2013) and Shi et al. (2013) on the ICDAR 2013 dataset. The proposed text detection and extraction methods achieved the best results in both recall and precision measures. The results are expressed as (recall, precision) value pair.: The proposed method obtained (81.04, 88.7) as compared to (72.01, 87.64); (66.45, 88.47); (64.84, 87.51); (68.24, 78.89) for Sung et al. (2015), Yin et al. (2014), Bai et al. (2013) and Shi et al. (2013) respectively.

Table 4
Multi-orientation extraction - evaluation results of the proposed text detection and extraction methods for ICDAR robust reading (2013) and KAIST datasets

Dataset	**ICDAR 2013**		**KAIST**	
Method	**Recall**	**Precision**	**Recall**	**Precision**
Sung et al. (2015)	72.01%	87.64%	-	-
Yin et al. (2014)	66.45%	88.47%	-	-
Bai et al. (2013)	64.84%	87.51%	-	-
Shi et al. (2013)	68.24%	78.89%	78%	66%
Kumar and Lee (2010)	-	-	60%	69%
Proposed method	81.04%	88.71%	83%	81%

As shown in Table 4, the proposed technique yielded a higher recall and precision values than any other state-of-the-art methods using the ICDAR dataset. Figures 5 and 6, which illustrate a sample of the extraction results on horizontal texts and non-horizontal text, respectively, demonstrate the effectiveness of the proposed method in detecting many text objects, including text with different colors, different lightings, complex backgrounds, different sizes, different stroke widths, flexible surfaces, other symbols, non-horizontal directions as well as low contrasts.

For the evaluation using the KAIST dataset, the proposed detection and extraction method outperformed Shi et al. (2013) and Kumar and Lee (2010) in extracting text from more complex and challenging scene images such as various font sizes, colors and abrupt environmental changes. The results in Table 4 shows the proposed detection and extraction methods achieved its best performance in both precision and recall whereby the proposed method achieved (0.81, 0.83) as compared to (0.66, 0.78) and (0.69, 0.60) for Shi et al. (2013) and Kumar and Lee (2010) respectively. From these results, it is observed that the proposed detection and extraction methods were able to extract text from the scene image with minimum negative error for different text width sizes and variants of text color. Furthermore, the proposed detection and extraction method also solved the problem of extracting characters for the exclusive case such as text art case outperforming other state-of-the-art methods (Figure 8). In summary, these results have been evaluated according to

the precision and recall metrics with respect to the proportions of the contents of the texts that have been extracted in comparison to the ground truth.

Despite the achievement, the strength and shortcoming of the state-of-the-art approaches are discussed below:

1. Sung et al. (2015) introduced a three-stage framework for character extraction, character verification and character refinement phases. Their character extraction mechanism consisted of ER tree construction, sub-path partitioning, sub-path pruning, and character candidate selection sequentially. Apart from employing AdaBoost trained character classifier for character verification, they commissioned heuristic rules for character and group refinement prior to geometric adjacency and color analysis. Regardless of its complexity and abandoning low contrast issue, the three-stage framework indicated significant results when dealing with multi-color and multi-oriented text.
2. Maximally Stable Extremal Regions (MSERs) algorithm initiated by Yin et al. (2014) was able to prune character candidates with the strategy of minimizing regularized variations. Furthermore, their single-link clustering algorithm so called a novel self-training distance metric learning algorithm was able to distinguish text and non-text areas automatically and efficiently. This innovation can be further accelerated if non-uniform illumination strategy is accounted.
3. Bai et al. (2014) proposed gradient local correlation for efficient scene text localization. In compliance to noise, small character size and shadow insensitivity, they developed a text confidence map by characterizing the density of pair wise edges and stroke width consistency and segmented text and non-text area using simple connected component, SVM classifier and color analysis.
4. By utilizing the available contrast of text pixels, Kumar and Lee (2010) claimed that their text extraction method was able to recover the missing text area. It comprised a set of algorithms such as pixel quantization, color to grey conversion, geometrical features, three layers of connected component and Sobel Projection which was able to recover missing text from the non-text region. Despite of its outstanding performance in dealing with non-uniform illumination, this handcrafted method could be further explored for achieving better results in non-uniform and oriented text.

For the text recognition evaluation using ICDAR 2013 and KAIST datasets, accuracy measure (Equation 8) which is the ratio of the number of correctly recognized words to the total number of words in the dataset is used. Accuracy is used instead of precision as the intention is just to demonstrate the performance of the proposed OCR method in recognizing the words in the dataset compare to the existing methods.

$$Accuracy = \frac{N_{tp} + N_{tn}}{N_{tp} + N_{fp} + N_{tp} + N_{tp}} \quad (8)$$

where N_{tp}, N_{fp}, N_{tn}, and N_{fn}, denote the number of True Positive, False Positive, True Negative and False Negative values, respectively. The evaluation results are tabulated in Table 5.

The results of the text recognition evaluation using ICDAR 2013 dataset in Table 5 show that the proposed OCR framework outperformed Novikova et al. (2012) method and was on par with Shi et al. (2013) method in terms of the number of words correctly recognized where the proposed method achieved 81.23% when compared to 82.1% and 57.99% for Shi et al. (2013) and Novikova et al. (2012) methods, respectively. Some sample results are highlighted in Figure 7.

For the KAIST dataset, results in Table 5 show that the proposed framework yielded better result than other methods in the recognition phase with the OCR accuracy of approximately 68% comparing to 56% and 63% for Text Spotter by Neumann and Matas (2015) and Gomez and Karatzas (2013), respectively. Furthermore, KAIST dataset contains categories of scene images according to the types of environments (outdoor, light and night, indoor and shadow). Table 6 demonstrates the effectiveness of the proposed OCR framework in the different types of environment. The proposed OCR method achieved high rate recognition in the outdoor and shadow environment types, 87% and 86.01% respectively.

Table 5
Text recognition evaluation results for ICDAR robust reading (2013) and KAIST datasets

Dataset	Method	Percentage of correctly recognized words
ICDAR Reading	Shi et al. (2013)	82.1%
	Novikova et al. (2012)	57.99%
	Proposed method	81.23%
KAIST	Text Spotter Neumann and Matas (2015)	56.4%
	Gomez and Karatzas (2013)	62.99%
	Proposed method	68.06%

Table 6
Results of the proposed OCR framework in KAIST dataset pertaining to four environments

Type	No of words	Recognized words	OCR accuracy
Outdoor	184	160	87.0
Light & night	38	27	71.1
Indoor	91	62	68.1
Shadow	36	31	86.01

The following are some benefits and limitations of the state-of-the-art word recognizers that may affect the experimental results:

Novikova et al. (2012) integrated local likelihood and pairwise positional consistency priors mainly for enforcing consistency of characters (lexicon) and their attributes in terms of font and color. Their word recognition process successfully estimated the maximum a posteriori or MAP inference under the joint posterior distribution of the model namely weighted finite-state transducers. Perhaps the mapping of font and color attributes in MAP inference requires further improvising.

Shi et al. (2013) applied part-based tree-structure to detect and recognize each type of character simultaneously. Their framework included modelling potential character locations using Conditional Random Field that incorporated detection scores, spatial constraints and linguistic knowledge. In the final stage, their word recognition result was achieved through the usage of cost function minimization in the random field.

Gomez and Karatzas (2013) also initiated a perceptual framework that exploited collaboration of proximity and similarity laws to create text-group hypotheses.

Driven by similar motivation, Neumann and Matas (2015) introduced a two-stage approach comprises a sequential selection against time from the set of Extremal Regions (ER) and clustering algorithm. They claimed ER was robust against blur, low contrast and illumination, color and texture variation because only ERs with locally maximal probability were nominated for the classification phase.

Overall, our proposed method has been shown to be effective in recognizing text of different forms i.e. size, color, stroke width, and orientation under different environmental condition i.e. outdoor, indoor, light and night, and shadow, outperforming the benchmarked state-of-the-art methods. Our method works especially well under the outdoor and shadow environments. Further work needs to be carried out to improve text recognition in the light and night as well as indoor categories. A robust text recognition method that is reasonably accurate in recognizing different fonts across a variety of environment is particularly attractive for applications that involve mobility such as detecting car license plate on vehicles on the highways as well as in indoor car park areas, providing scene information to support the visually impaired people via scene-to-text-to-speech translation as they carry out their daily activities in various environments and recognition.

CONCLUSION

In this paper, a novel text detection/extraction and recognition technique for real scene text images that are tolerant of different types of degradation is presented. The main objective of the proposed detection and extraction method and OCR framework is to tackle the extraction issues and difficulties faced when dealing with natural scene text images, specifically in solving multi-oriented text, the lighting effect on the appearance of text, and

low contrast. The proposed text extraction and detection method consists of three steps: first is the exploited NTSC enhancement, second is the hybrid adaptive segmentation combining fuzzy C-means segmentation and adaptive binarization, and third is a rule-based text extraction with stroke width transform using statistical and geometrical features during text filtering for detecting connected component regions. After distinguishing text from non-text extracted from a scene image, text recognition is accomplished using the Tesseract OCR engine.

To evaluate the proposed method, it was compared with the state-of-the-art detection/extraction and recognition methods. All the experiments were conducted using the ICDAR 2013 robust reading dataset and KAIST scene text dataset. Experimentally, it is established that the proposed method outperformed current state-of-the-art extraction methods, achieving >80% in both precision and recall measures. This is because the proposed method enhances the quality of the scene text image source by solving most of the text detection problems such as high-orientation text, low contrast, and illumination for better text extraction and recognition.

ACKNOWLEDGEMENTS

The authors would like to thank the Faculty of Information Science and Technology and the Center for Artificial Intelligence Technology, Universiti Kebangsaan Malaysia for providing the facilities and financial support under the grant: AP2017-05/2 and FRGS-1-2019-ICT02-UKM-02-9.

REFERENCES

Bai, B., Yin, F., & Liu, C. L. (2013, August 25-28). Scene text localization using gradient local correlation. In *2013 12th International Conference on Document Analysis and Recognition (ICDAR)* (pp. 1380-1384). Washington, DC, USA.

Behzadi, M., & Safabakhsh, R. (2018, December 25-27). Text Detection in Natural Scenes using Fully Convolutional DenseNets. In *2018 4th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)* (pp. 11-14). Tehran, Iran.

Burney, S. A., & Tariq, H. (2014). K-means cluster analysis for image segmentation. *International Journal of Computer Applications*, *96*(4), 1-8.

Chen, H., Tsai, S. S., Schroth, G., Chen, D. M., Grzeszczuk, R., & Girod, B. (2011, September 11-14). Robust text detection in natural images with edge-enhanced maximally stable extremal regions. In *2011 18th IEEE International Conference on Image Processing (ICIP),* (pp. 2609-2612). Brussels, Belgium.

Gomez, L., & Karatzas, D. (2013, August 25-28). Multi-script text extraction from natural scenes. In *2013 12th International Conference on Document Analysis and Recognition* (pp. 467-471). Washington, DC, USA.

Horvath, J. (2006, January 20-21). Image segmentation using fuzzy c-means. In *Proceedings of SAMI, 4th Slovakian-Hungarian Joint Symposium on Applied Machine Intelligence* (pp. 144-151). Herlany, Slovakia.

Huang, S., Xu, H., Xia, X., & Zhang, Y. (2018, December 8-9). End-to-end vessel plate number detection and recognition using deep convolutional neural networks and LSTMs. In *2018 11th International Symposium on Computational Intelligence and Design (ISCID)* (Vol. 1, pp. 195-199). Hangzhou, China.

Jadhav, A. J., Kolhe, V., & Peshwe, S. (2013). Text Extraction from Images: a survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, *3*(3), 333-337.

Jung, K., Kim, K. I., & Jain, A. K. (2004). Text information extraction in images and video: a survey. *Pattern Recognition*, *37*(5), 977-997.

Karanje, U. B., & Dagade, R. (2014). Survey on Text Detection, Segmentation and Recognition from a Natural Scene Images. *International Journal of Computer Applications*, *108*(13), 39-43.

Kraisin, S., & Kaothanthong, N. (2018, November 15-17). Accuracy Improvement of a Province Name Recognition on Thai License Plate. In *2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)* (pp. 1-6). Pattaya, Thailand.

Kumar, M., & Lee, G. (2010, February 26-28). Automatic text location from complex natural scene images. In *2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE)* (Vol. 3, pp. 594-597). Singapore.

Kumar, M., Kim, Y. C., & Lee, G. S. (2010, June 29-July 1). Text detection using multilayer separation in real scene images. In *2010 IEEE 10th International Conference on Computer and Information Technology (CIT)* (pp. 1413-1417). Bradford, UK.

Li, H., Wang, P., & Shen, C. (2018). Toward end-to-end car license plate detection and recognition with deep neural networks. *IEEE Transactions on Intelligent Transportation Systems, 20*(3), 1126-1136.

Nagaraju, G., Ramaraju, P. V., Sandeep, P. M. M. P., Nawaz, S. M., & Bhargav, S. K. (2015). Text Extraction from Images with Edge-Enhanced Mser and Hardware Interfacing Using Arduino. *International Journal of Engineering and Computer Science*, *4*(05), 11797-11803.

Neumann, L., & Matas, J. (2015). Real-time lexicon-free scene text localization and recognition. *IEEE transactions on pattern analysis and machine intelligence*, *38*(9), 1872-1885.

Ning, G., Han, T. X., & He, Z. (2015, September 27-30). Scene text detection based on component-level fusion and region-level verification. In *2015 IEEE International Conference on Image Processing (ICIP)* (pp. 837-841). Quebec City, QC, Canada.

Nosal, E. M. (2008, October 14-17). Flood-fill algorithms used for passive acoustic detection and tracking. In *2008 New Trends for Environmental Monitoring Using Passive Systems* (pp. 1-5). Hyeres, French.

Novikova, T., Barinova, O., Kohli, P., & Lempitsky, V. (2012). Large-lexicon attribute-consistent text recognition in natural images. In *European Conference on Computer Vision* (pp. 752-765). Heidelberg, Germany: Springer.

Phan, T. Q., Shivakumara, P., & Tan, C. L. (2009, July 26-29). A Laplacian method for video text detection. In *2009 10th International Conference on Document Analysis and Recognition* (pp. 66-70). Barcelona, Spain.

Qaisar, S. M., Khan, R., & Hammad, N. (2019, March 26-April 10). Scene to Text Conversion and Pronunciation for Visually Impaired People. In *2019 Advances in Science and Engineering Technology International Conferences (ASET)* (pp. 1-4). Dubai, United Arab Emirates.

Raghunandan, K. S., Shivakumara, P., Roy, S., Kumar, G. H., Pal, U., & Lu, T. (2018). Multi-script-oriented text detection and recognition in video/scene/born digital images. *IEEE Transactions on Circuits and Systems for Video Technology, 29*(4), 1145-1162.

Raj, H., & Ghosh, R. (2014, September 24-27). Devanagari text extraction from natural scene images. In *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 513-517). New Delhi, India.

Rajaby, E., Ahadi, S. M., & Aghaeinia, H. (2016). Robust color image segmentation using fuzzy c-means with weighted hue and intensity. *Digital Signal Processing*, *51*, 170-183.

Risnumawan, A., Shivakumara, P., Chan, C. S., & Tan, C. L. (2014). A robust arbitrary text detection system for natural scene images. *Expert Systems with Applications*, *41*(18), 8027-8048.

Samadhiya, A., & Khatri, P. (2014). General Review of Various Text. *International Journal of Advances in Electronics and Computer Science, 2*(1), 12-16.

Shah, K. M. & Thakar, V. K. (2010). Content Based Image Retrieval using different Colormaps. In C. Modi, H. Shah, R. Kher & N. Desai (Eds.), *Proceedings of the 2009 International Conference on Signals, Systems and Automation (ICSSA 2009)*. Boca Raton, Florida: Universal-Publishers.

Shi, C., Wang, C., Xiao, B., Zhang, Y., Gao, S., & Zhang, Z. (2013, June 23-28). Scene text recognition using part-based tree-structured character detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2961-2968). Portland, Oregon, USA.

Shivakumara, P., Phan, T. Q., & Tan, C. L. (2011). A laplacian approach to multi-oriented text detection in video. *IEEE transactions on pattern analysis and machine intelligence*, *33*(2), 412-419.

Sulaiman, A., Omar, K., & Nasrudin, M. F. (2019). Degraded historical document binarization: A review on issues, challenges, techniques, and future directions. *Journal of Imaging, 5*(4), 48-73. doi: 10.3390/jimaging5040048

Sumathi, C. P., Santhanam, T., & Devi, G. G. (2012). A survey on various approaches of text extraction in images. *International Journal of Computer Science and Engineering Survey*, *3*(4), 27-42.

Sun, Y., Mao, X., Hong, S., Xu, W., & Gui, G. (2019). Template Matching-Based Method for Intelligent Invoice Information Identification. *IEEE Access, 7*, 28392-28401.

Sung, M. C., Jun, B., Cho, H., & Kim, D. (2015, August 23-26). Scene text detection with robust character candidate extraction method. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)* (pp. 426-430). Tunis, Tunisia.

Wang, L., Huang, L. L., & Wu, Y. (2011, November 28). An efficient coarse-to-fine scheme for text detection in videos. In *2011 First Asian Conference on Pattern Recognition (ACPR)* (pp. 475-479). Beijing, China.

Wang, Q., Lu, Y., & Sun, S. (2015, August 23-26). Text detection in nature scene images using two-stage nontext filtering. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)* (pp. 106-110). Tunis, Tunisia.

Yao, J. L., Wang, Y. Q., Weng, L. B., & Yang, Y. P. (2007, November 2-4). Locating text based on connected component and SVM. In *2007International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR'07)* (Vol. 3, pp. 1418-1423). Beijing, China.

Yin, X. C., Yin, X., Huang, K., & Hao, H. W. (2014). Robust text detection in natural scene images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(5), 970-983.

Zhang, H., Zhao, K., Song, Y. Z., & Guo, J. (2013). Text extraction from natural scene image: A survey. *Neurocomputing*, *122*, 310-323.