Pattern Recognition 42 (2009) 1419 – 1444

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

A multi-plane approach for text segmentation of complex document images

Yen-Lin Chen[a], Bing-Fei Wu[b,*]

[a]Department of Computer Science and Information Engineering, Asia University, 500 Liufeng Road, Wufeng, Taichung 41354, Taiwan
[b]Department of Electrical and Control Engineering, National Chiao Tung University, 1001 Ta-Hsueh Road, Hsinchu 30010, Taiwan

ARTICLE INFO

Article history:
Received 19 January 2008
Received in revised form 1 September 2008
Accepted 19 October 2008

Keywords:
Document image processing
Text extraction
Image segmentation
Multilevel thresholding
Region segmentation
Complex document images

ABSTRACT

This study presents a new method, namely the multi-plane segmentation approach, for segmenting and extracting textual objects from various real-life complex document images. The proposed multi-plane segmentation approach first decomposes the document image into distinct object planes to extract and separate homogeneous objects including textual regions of interest, non-text objects such as graphics and pictures, and background textures. This process consists of two stages—localized histogram multilevel thresholding and multi-plane region matching and assembling. Then a text extraction procedure is applied on the resultant planes to detect and extract textual objects with different characteristics in the respective planes. The proposed approach processes document images regionally and adaptively according to their respective local features. Hence detailed characteristics of the extracted textual objects, particularly small characters with thin strokes, as well as gradational illuminations of characters, can be well-preserved. Moreover, this way also allows background objects with uneven, gradational, and sharp variations in contrast, illumination, and texture to be handled easily and well. Experimental results on real-life complex document images demonstrate that the proposed approach is effective in extracting textual objects with various illuminations, sizes, and font styles from various types of complex document images.

1. Introduction

Extraction of textual information from document images provides many useful applications in document analysis and understanding, such as optical character recognition, document retrieval, and compression [1,2]. To-date, many techniques were presented for extracting textual objects from monochromatic document images [3–6]. In recent years, advances in multimedia publishing and printing technology have led to an increasing number of real-life documents in which stylistic character strings are printed with pictorial, textured, and decorated objects and colorful, varied background components. However, most of current approaches cannot work well for extracting textual objects from real-life complex document images. Compared to monochromatic document images, text extraction in complex document images brings many difficulties associated with the complexity of background images, variety, and shading of character illuminations, the superimposing of characters with illustrations and pictures, as well as other decorated background components. As a result, there is an increasing demand for a system that is able to read and extract the textual information printed on pictorial and textured regions in both colored images as well as monochromatic main text regions.

Several newly developed global thresholding methods are useful in separating textual objects from non-uniform illuminated document images. Liu and Srihari [7] proposed a method based on texture features of character patterns, while Cheriet et al. [8] presented a recursive thresholding algorithm extended from Otsu's optimal criterion [9]. These methods are performed by classifying pixels in the original image as foreground objects (particularly textual objects of interest) or as background ones according to their gray intensities in a global view, and are attractive because of computational simplicity. However, binary images obtained by global thresholding techniques are subject to noise and distortion, especially because of uneven illumination and the spreading effect caused by the image scanner. To solve the above-mentioned issues, Solihin and Leedham's integral ratio approaches [10] provided a new class of histogram-based thresholding techniques which classify pixels into three classes: foreground, background, and a fuzzy region between two basic classes. In Ref. [11], Parker proposed a local gray intensity gradient thresholding technique which is effective for extracting textual objects in badly illuminated document images. Because this method is based on the assumption of binary document images, its application is limited to extracting character objects from backgrounds no more complex than monotonically changing illuminations. A local and adaptive

* Corresponding author. Tel.: +886 3 5131538; fax: +886 3 5712385.
E-mail addresses: ylchen@asia.edu.tw (Y.-L. Chen), bwu@cssp.cn.nctu.edu.tw (B.-F. Wu).

0031-3203/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.10.032

binarization method was presented by Ohya et al. [12]. This method divides the original image into blocks of specific size, determines an optimal threshold associated with each block to be applied on its center pixel, and uses interpolation for determining pixel-wise thresholds. It can effectively extract textual objects from images with complex backgrounds on condition that the illuminations are very bright compared with those of the textual objects.

Some other methods support a different viewpoint for extracting texts by modeling the features of textual objects and backgrounds. Kamel and Zhao [13] proposed the logical level technique to utilize local linearity features of character strokes, while Venkateswarlu and Boyle's average clustering algorithm [14] utilizes local statistical features of textual objects. These methods apply symmetric local windows with a pre-specified size, and several pre-determined thresholds of prior knowledge on the local features, and so that characters with stroke widths that are substantially thinner or thicker than the assumed stroke width, or characters in varying illumination contrasts with backgrounds may not be appropriately extracted. To deal with these problems, Yang and Yan [15] presented an adaptive logical method (ALM) which applies the concepts of Liu and Srihari's run-length histogram [7] on sectored image regions, to provide an effective scheme for automatically adjusting the size of the local window and logical thresholding level. Ye et al.'s hybrid extraction method [16] integrates global thresholding, local thresholding, and the double-edge stroke feature extraction techniques to extract textual objects from document images with different complexities. The double-edge technique is useful in separating characters whose stroke widths are within a specified size from uneven backgrounds. Some recently presented methods [17,18] utilized the sub-image concepts to deal with the extraction of textual objects under different illumination contrasts with backgrounds. Dawoud and Kamel's [17] proposed a multi-model sub-image thresholding method that considers a document image as a collection of predetermined regions, i.e. sub-images, and then textual objects contained in each sub-image are segmented using statistical models of the gray-intensity and stroke-run features. In Amin and Wu's multistage thresholding approach [18] Otsu's global thresholding method is firstly applied, and then a connected-component labeling process is applied on the thresholded image to determine the sub-images of interest, and these sub-images then undergo another thresholding process to extract textual objects. The extraction performance of the above two methods relies principally on the adequate determination of sub-image regions. Thus, in case of the textual objects overlapping on pictorial or textured backgrounds of poor and varying contrasts, suitable sub-images are hard to determine to obtain satisfactory extraction results.

Since most textual objects show sharp and distinctive edge features, methods based on edge information [19–22] have been developed. Such methods utilize an edge detection operator to extract the edge features of textual objects, and then use these features to extract texts from document images. Wu et al.'s textfinder system [20] uses nine second-order Gaussian derivative filters to obtain edge-feature vectors of each pixel at three different scales, and applies the *K*-means algorithm on these edge-feature vectors to identify corresponding textual pixels. Hasan and Karam [21] introduced a method that utilizes a morphological edge extraction scheme, and applies morphological dilation and erosion operations on the extracted closure edges to locate textual regions. Edge information can also be treated as a measure for detecting the existence of textual objects in a specific region. In Pietikainen and Okun's work [22], edge features extracted by the Sobel operator are divided into non-overlapping blocks, and then these blocks are classified as text or non-text according to their corresponding values of the edge features. Such edge-based methods are capable of extracting textual objects in different homogeneous illuminations from graphic backgrounds. However, when the textual objects are adjoined or touched with graphical objects, texture patterns, or backgrounds with sharply varying contours, edge-feature vectors of non-text objects with similar characteristics may also be identified as textual ones, and thus the characters in extracted textual regions are blurred by those non-text objects. Moreover, when textual objects do not have sufficient contrasts with non-text objects or backgrounds to form sufficiently strong edge features, such textual objects cannot be easily extracted with edge-based methods.

In recent years, several color-segmentation-based methods for text extraction from color document images have been proposed. Zhong et al. [23] proposed two methods and a hybrid approach for locating texts in color images, such as in CD jackets and book covers. The first method utilizes a histogram-based color clustering process to obtain connected-components with uniform colors, and then several heuristic rules are applied to classify them as textual or non-textual objects. The second method locates textual regions based on their distinctive spatial variance. To detect textual regions more effectively, both methods are combined into a hybrid approach. Although the spatial variance method still suffers from the drawbacks of the edge-based methods mentioned previously, the color connected-component method moderately compensates for these drawbacks. However, this approach still cannot provide acceptable results when the illuminations or colors of characters in large textual regions are shaded. Several recent techniques utilize color clustering or quantization approaches to determine the prototype colors of documents so as to facilitate the detection of character objects in these separated color planes. In Jain and Yu's work [24], a color document is decomposed into a set of foreground images in the RGB color space using a bit-dropping quantization and the single-link color clustering algorithm. Strouthopoulos et al.'s adaptive color reduction technique [25] utilizes an unsupervised neural network classifier and a tree-search procedure to determine prototype colors. Some alternative color spaces are also adopted to determine prototype colors for finding textual objects of interest. Yang and Ozawa [26] make use of the HSI color space to segment homogenous color regions to extract bibliographic information from book covers, while Hase et al. [27] apply a histogram-based approach to select prototype colors on the CIE *Lab* color space to obtain textual regions. However, most of the aforementioned methods have difficulties in extracting texts which are embedded in complex backgrounds or that touch other pictorial and graphical objects. This is because the prototype colors are determined in a global view, so that appropriate prototype colors cannot be easily selected to distinguish textual objects from those touched pictorial objects and complex backgrounds without sufficient contrasts. Furthermore, such problems also limit the reliability of such methods in handling unevenly illuminated document images.

In brief, extracting texts from complex document images involves several difficulties. These difficulties arise from the following properties of complex documents: (1) character strings in complex document images may have different illuminations, sizes, font styles, and may be overlapped with various background objects with uneven, gradational, and sharp variations in contrast, illumination, and texture, such as illustrations, photographs, pictures or other background textures and (2) these documents may comprise small characters with very thin strokes as well as large characters with thick strokes, and may be influenced by image shading. An approach for extracting black texts from such complex backgrounds to facilitate compression of document images has been proposed in our previous work [28].

In this study, we propose an effective method, namely the *multi-plane segmentation approach*, for segmenting and extracting textual objects of interest from these complex document images, and resolving the above issues associated with the complexity of their backgrounds. The proposed multi-plane segmentation approach first

decomposes the document image into distinct object planes to extract and separate homogeneous objects including textual regions of interest, non-text objects such as graphics and pictures, and background textures. This process consists of two stages—*localized histogram multilevel thresholding* and *multi-plane region matching and assembling*. Then a text extraction procedure is applied on the resultant planes to detect and extract textual objects with different characteristics in the respective planes. The proposed approach processes document images regionally and adaptively by means of their local features. This way allows detailed characteristics of the extracted textual objects to be well-preserved, especially the small characters with thin strokes, as well as characters in gradational and shaded illumination contrasts. Thus, textual objects adjoined or touched with pictorial objects and backgrounds with uneven, gradational, and sharp variations in contrast, illumination, and texture can be handled easily and well. Experimental results demonstrate that the proposed approach is capable of extracting textual objects with different illuminations, sizes, and font styles from different types of complex document images. As compared with other existing techniques, our proposed approach exhibits feasible and effective performance on text extraction from various real-life complex document images.

2. Overview of the proposed approach

The proposed multi-plane segmentation approach decomposes the document image into separate object planes by applying the two processing stages: automatic localized histogram multilevel thresholding, and multi-plane region matching and assembling. The flow diagram of the proposed approach is illustrated in Fig. 1. In the first stage, the original image is firstly sectored into non-overlapping "localized block regions", denoted by $\Re^{ij}$, then distinct objects embedded in block regions are decomposed into separate "sub-block regions (*SRs*)" by applying the localized histogram multilevel thresholding process, as illustrated in Figs. 2–4. Afterward, in the second stage, the multi-plane region matching and assembling process, which adopts both the localized spatial dissimilarity relation and the global feature information, is applied to perceptually classify and arrange the obtained *SRs* to compose a set of homogeneous "object planes", denoted by $\mathscr{P}_q$, especially textual regions of interest. This proposed multi-plane region matching and assembling process is conducted by recursively applying the following three phases—the initial plane selection phase, the matching phase, and the plane construction phase, as depicted in Fig. 6. Consequently, homogeneous objects including textual regions of interest, non-text objects such as graphics and pictures, and background textures are extracted and separated into distinct object planes. The text extraction process is then performed on the resultant planes to extract the textual objects with different characteristics in the respective planes, as shown in Fig. 7. The important symbols utilized for the presentation of the proposed approach are depicted in Table 1.

The following sections will accordingly describe the detailed stages of the proposed approach, and are organized as follows. In Sections 3 and 4, the two stages of the proposed multi-plane segmentation approach, the localized histogram multilevel thresholding procedure, and the multi-plane region matching and assembling process, are, respectively, presented. Then, a simple text extraction procedure is described in Section 5. Next, Section 6 illustrates parameter adaptation and comparative performance evaluation results. Finally, the conclusions of this study are stated in Section 7.

3. Localized histogram multilevel thresholding

For complex document images with textual objects in different illuminations, sizes, and font styles, and printed on varying or

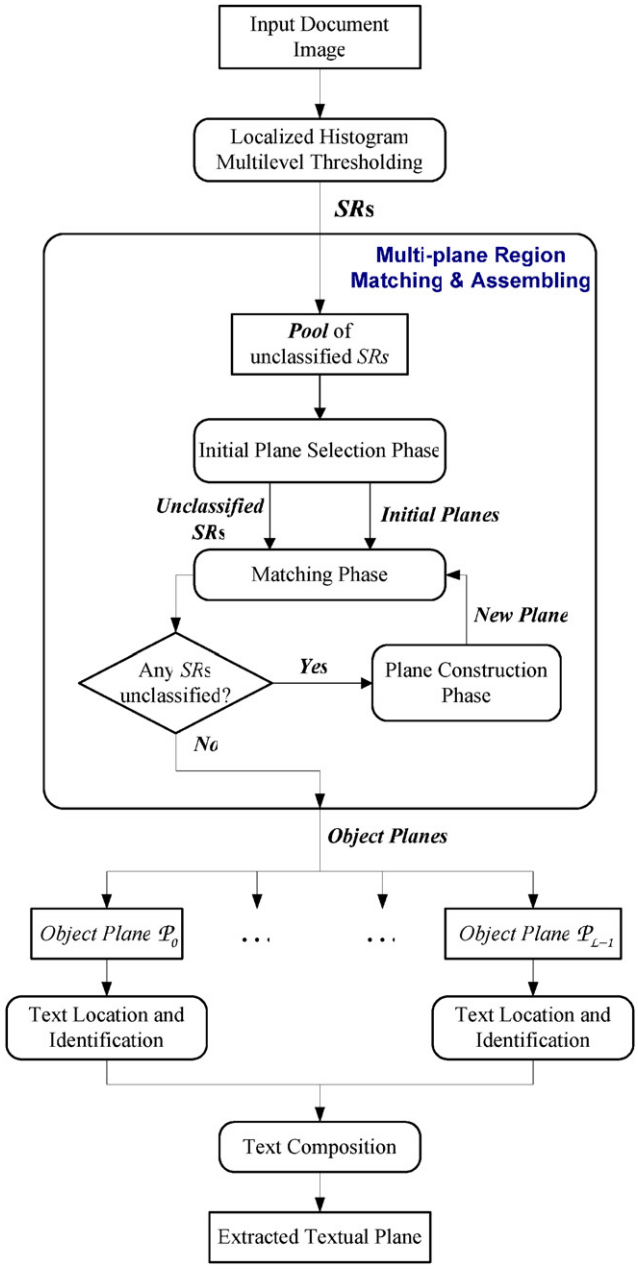

Fig. 1. Block diagram of the proposed multi-plane segmentation approach.

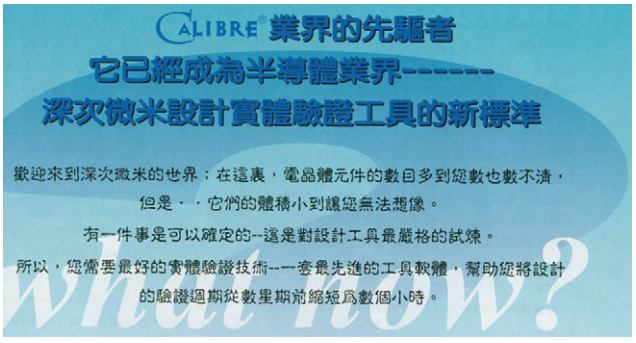

Fig. 2. A 1929×1019 original complex document image.

inhomogeneous background objects with uneven, gradational, and sharp variations in contrast, illumination, and texture, such as illustrations, photographs, pictures or other background patterns, a critical difficulty arises that no global segmentation techniques could

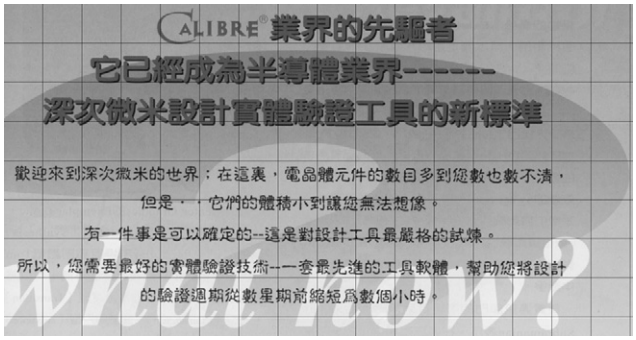

Fig. 3. Sectored regions of the illumination image **Y** obtained from the original image in Fig. 2.

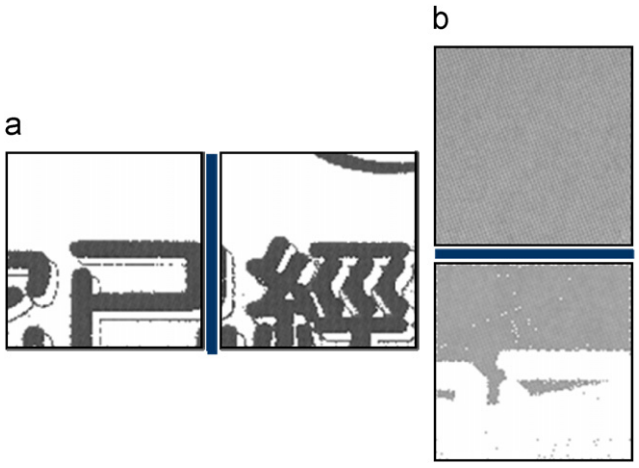

Fig. 5. Types of touching boundaries of the two 4-adjacent *SR*s: (a) vertical boundary and (b) horizontal boundary.

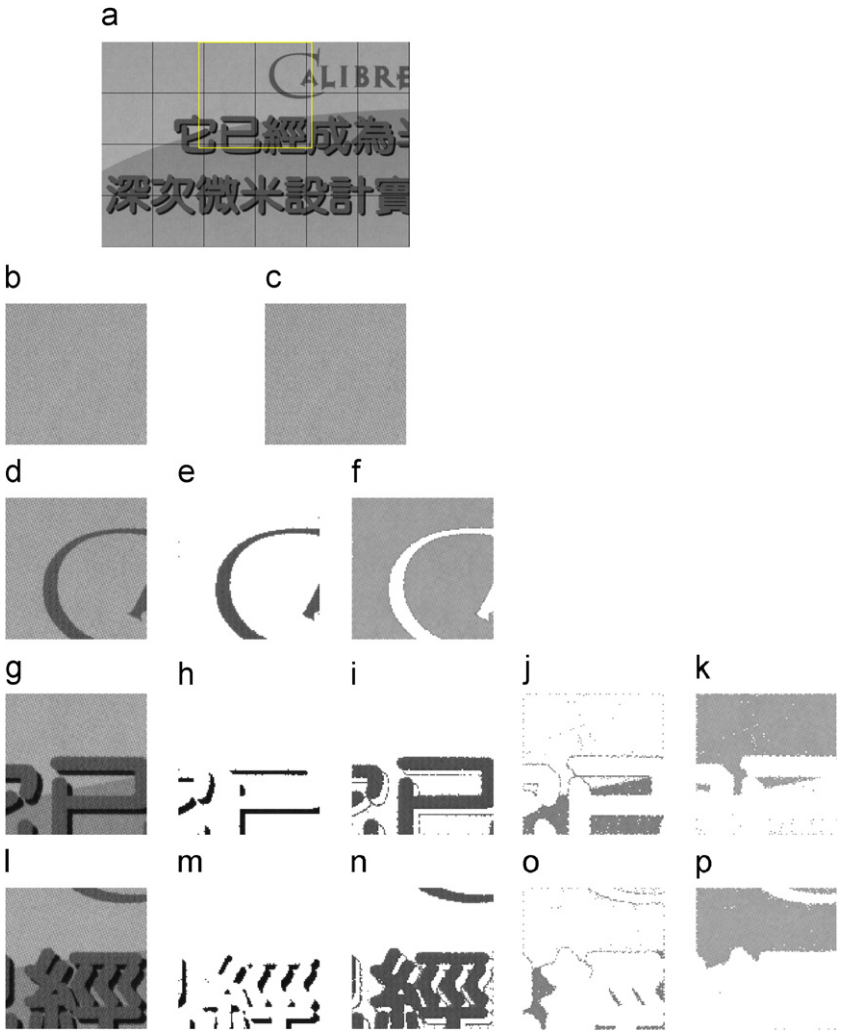

Fig. 4. Example of the results by the localized multilevel thresholding procedure, and the resultant $\mathscr{SF}$ values of $\Re^{i_2 j_1}$, $\Re^{i_1 j_2}$, and $\Re^{i_2 j_2}$ after the thresholding procedure are 0.931, 0.961, and 0.96, respectively: (a) part of the partitioned block regions of the image "Calibre" in Fig. 3, where the block regions enclosed by yellow ink are employed for the following examples of the localized multilevel thresholding procedure, (b) the upper-left block region, $\Re^{i_1 j_1}$, $\mathscr{SF}_b = 0.577$, and $\sigma_{\Re} = 8.81$, (c) $SR^{i_1 j_1,0}$ derived from $\Re^{i_1 j_1}$, which is a homogenous block region, (d) the upper-right block region $\Re^{i_2 j_1}$, $\mathscr{SF}_b = 0.931$, and $\sigma_{\Re} = 22.8$, (e) $SR^{i_2 j_1,0}$ derived from $\Re^{i_2 j_1}$, (f) $SR^{i_2 j_1,1}$ derived from $\Re^{i_2 j_1}$, (g) the bottom-left block region $\Re^{i_1 j_2}$, $\mathscr{SF}_b = 0.804$, and $\sigma_{\Re} = 42.3$, (h) $SR^{i_1 j_2,0}$ derived from $\Re^{i_1 j_2}$, (i) $SR^{i_1 j_2,1}$ derived from $\Re^{i_1 j_2}$, (j) $SR^{i_1 j_2,2}$ derived from $\Re^{i_1 j_2}$, (k) $SR^{i_1 j_2,3}$ derived from $\Re^{i_1 j_2}$, (l) bottom-right block region $\Re^{i_2 j_2}$, $\mathscr{SF}_b = 0.835$, and $\sigma_{\Re} = 46.6$, (m) $SR^{i_2 j_2,0}$ derived from $\Re^{i_2 j_2}$, (n) $SR^{i_2 j_2,1}$ derived from $\Re^{i_2 j_2}$, (o) $SR^{i_2 j_2,2}$ derived from $\Re^{i_2 j_2}$, and (p) $SR^{i_2 j_2,3}$ derived from $\Re^{i_2 j_2}$.

Fig. 6. An example of the test image, "Calibre", and the object planes obtained by the multi-plane segmentation (image size = 1929×1019): (a) object plane $\mathscr{P}_0$, (b) object plane $\mathscr{P}_1$, (c) object plane $\mathscr{P}_2$, (d) object plane $\mathscr{P}_3$, (e) object plane $\mathscr{P}_4$, (f) object plane $\mathscr{P}_5$, and (g) object plane $\mathscr{P}_6$.

work well for such kinds of document images. This is because when the regions of interesting textual objects consisted of multiple colors or gray intensities are undersized as compared with those of the touched pictorial objects and complex backgrounds with indistinct contrasts, these textual objects cannot be discriminated in a global view of statistical features. A typical example with these characteristics is shown in Fig. 2. This sample image consists of three different colored textual regions printed on a varying and shaded background. Moreover, the black characters are superimposed on the white characters. By observing some localized regions, the statistical features of the textual objects, pictorial objects, and backgrounds could be much more distinguishable. Therefore, regional and adaptive analysis approach for the localized statistical features can provide detailed characteristics of the textual objects of interest to be well-extracted for later document processing. In this section, we will introduce a simple and effective localized segmentation approach as the first stage of the multi-plane segmentation process for extracting textual objects from complex document images.

The multi-plane segmentation process, if necessary, begins by applying a color-to-grayscale transformation on the *RGB* components of image pixels in a color document image, to obtain its illumination image **Y**. After the color transformation is performed, the illumination image **Y** still retain the texture features of the original color image, as pointed out in Ref. [20], and thus the character strokes in their original color are still well-preserved. Then the obtained illumination image **Y** will be sectored into non-overlapping localized block regions $\Re^{i,j}$ with a given size $M_H \times M_V$, as shown in Fig. 3. To facilitate analysis in the following stage, the objects of interest must be extracted from these localized block regions into separate *SR*s, each of which contains objects with homogeneous features. Toward this goal, the discriminant criterion is useful for measuring separability among the decomposed regions with different objects. Its application on bi-level global thresholding to extract foreground objects

from the background was first presented by Otsu [9]. This method is ranked as the most effective bi-level threshold selection method [29,30]. However, when the number of desired thresholds increases, the computation needed to obtain the optimal threshold values is substantially increased and the search to achieve the optimal value of the criterion function is particularly exhaustive.

Hence, an efficient multilevel thresholding technique is needed to automatically determine the suitable number of thresholds to segment the block region into different decomposed object regions. By using the properties of discriminant analysis, we have proposed an automatic multilevel global thresholding technique for image segmentation [31]. This technique extends and applies the concept of discriminant criterion on analyzing the separability among the gray levels in the image. It can automatically determine the suitable number of thresholds, and utilizes a fast recursive selection strategy to select the optimal thresholds to segment the image into separate objects with similar features in a computationally frugal way. Based on this effective technique, we will introduce a localized histogram multilevel thresholding process to decompose distinct objects with homogeneous features in localized block regions into separate *SRs*. This process is described in the following subsections.

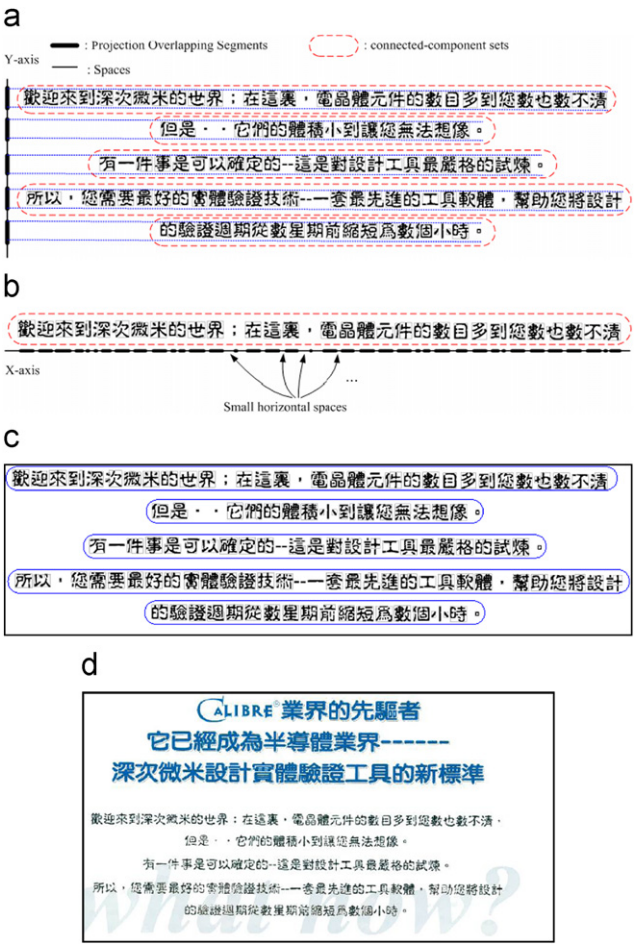

Fig. 7. Examples of the text location and extraction process: (a) example of performing *X-cut* on connected-components in the binary plane $\mathscr{BP}_4$ of Fig. 6(g), (b) example of performing *Y-cut* on the top connected-component group, which is the first group among five groups obtained from *X-cut* procedure on $\mathscr{BP}_4$, (c) the resultant candidate text-lines obtained by the *XY*-cut spatial clustering process, and (d) the resultant text plane obtained by performing text extraction process on all object planes derived from Fig. 2.

3.1. Statistical features and recursive partition concepts of localized regions

Let f_g denote the observed frequencies (histogram) of gray intensities of pixels in a localized block region $\Re^{i,j}$ with a given gray intensity g, and thus the total amount of pixels in $\Re^{i,j}$ can be given by $N = f_0 + f_1 + \ldots + f_{U-1}$, where U is the number of gray intensities in the histogram. Hence, the normalized probability of one pixel having a given gray intensity can be computed as,

$$P_g = \frac{f_g}{N}, \quad \text{where } P_g \geqslant 0, \quad \text{and} \quad \sum_{g=0}^{U-1} P_g = 1 \tag{1}$$

In order to segment textual objects, foreground objects and background components from a given localized region $\Re^{i,j}$, pixels in $\Re^{i,j}$ should be partitioned into a suitable number of classes. For multilevel thresholding, with n thresholds to partition the pixels in the region $\Re^{i,j}$ into $n+1$ classes, gray intensities of pixels in $\Re^{i,j}$ are segmented by applying a threshold set $\boldsymbol{T}$, which is composed of n thresholds, where $\boldsymbol{T} = \{t_k | k = 1, \ldots, n\}$. These classes are represented by $C_0 = \{0, 1, \ldots, t_1\}, \ldots C_k = \{t_k + 1, t_k + 2, \ldots, t_{k+1}\}, \ldots, C_n = \{t_n + 1, t_n + 2, \ldots, U-1\}$. Then the statistical features associated with a given pixel class C_k, including the cumulative probability, the mean, and the standard deviation, denoted by w_k, μ_k, and σ_k^2, respectively, can be computed as

$$w_k = \sum_{g=t_k+1}^{t_{k+1}} P_g, \quad \mu_k = \frac{\sum_{g=t_k+1}^{t_{k+1}} g P_g}{w_k}, \quad \text{and}$$
$$\sigma_k^2 = \frac{\sum_{g=t_k+1}^{t_{k+1}} P_g (g - \mu_k)^2}{w_k} \tag{2}$$

Based on the above-mentioned statistical features of pixels in the region $\Re^{i,j}$, the between-class variance, denoted by v_{BC}, an effective criterion for evaluating segmentation results, can be obtained for measuring the separability among all classes, and is expressed as

$$v_{BC}(\boldsymbol{T}) = \sum_{k=0}^{n} w_k (\mu_k - \mu_{\Re})^2, \quad \text{where } \mu_{\Re} = \sum_{g=0}^{U-1} g P_g \tag{3}$$

where $\mu_{\Re}$ is the overall mean of the gray intensities in $\Re^{i,j}$. Then the within-class variance and total variance, denoted by v_{WC} and

Table 1
List of important symbols of the proposed approach.

Symbol	Description
$\Re^{i,j}$	Localized block region, which represents one of the non-overlapping block regions sectored from the original image, and the superscript (i,j) denotes its location index
$SR^{i,j,k}$, $SR_q^{i',j',k'}$	Sub-block region, which is derived from $\Re^{i,j}$ after applying the localized histogram multilevel thresholding process; the additional superscript k means it is k-th *SRs* derived from $\Re^{i,j}$, and when the subscript q is assigned, it means that this *SR* has belonged to an existent object plane $\mathscr{P}_q$
$\mathscr{P}_q$	Object plane, which is formed by a set of homogeneous *SRs* after performing the multi-plane region matching and assembling process, and the subscript q represents its order of creation

$\sigma_{\Re}^2$, respectively, of all segmented classes of gray intensities are, respectively, computed as

$$\nu_{WC}(\boldsymbol{T}) = \sum_{k=0}^{n} w_k \sigma_k^2, \quad \sigma_{\Re}^2 = \sum_{g=0}^{U-1} (g - \mu_{\Re})^2 P_g \tag{4}$$

Here, a dummy threshold $t_0 = 0$ is utilized for the sake of convenience in simplifying the expression of equation terms.

The aforementioned criterion functions can be considered as a measure of separability among all existing classes decomposed from the original region $\Re^{i,j}$. We utilize this concept as a criterion of automatic segmentation of objects in a region, denoted by the "separability factor"—$\mathscr{SF}$ in this study, which is defined as

$$\mathscr{SF} = \frac{\nu_{BC}(\boldsymbol{T})}{\sigma_{\Re}^2} = 1 - \frac{\nu_{WC}(\boldsymbol{T})}{\sigma_{\Re}^2} \tag{5}$$

where $\sigma_{\Re}^2$ serves as the normalization factor in this equation. The $\mathscr{SF}$ value represents the separability measure among all existing classes, and lies within the range $\mathscr{SF} \in [0,1]$; the lower bound is approached when the region $\Re^{i,j}$ comprises a uniform gray intensity, while the upper bound is achieved when the region $\Re^{i,j}$ consists of exactly n+1 gray intensities. The objective is to maximize the $\mathscr{SF}$ value so as to optimize the segmentation result. This concept is supported by the property that $\sigma_{\Re}^2$ is equivalent to the sum of ν_{BC} and ν_{WC}. By observing the terms comprising $\nu_{WC}(\boldsymbol{T})$, if the gray intensities of the pixels belonging to most existing classes are widely distributed, i.e. the contribution values of their class variances σ_k^2 are large, then the value of the corresponding $\mathscr{SF}$ measure becomes low. Accordingly, when $\mathscr{SF}$ approximates 1.0, all resultant classes of gray intensities C_k ($k = 0, \ldots, n$), which are decomposed from the original region $\Re^{i,j}$, are ideally and completely separated.

Therefore, based on this efficient discriminant criterion, an automatic multilevel thresholding can be applied for recursively segmenting the block region $\Re^{i,j}$ into different objects of homogeneous illuminations, regardless of the number of objects and image complexity of the region $\Re^{i,j}$. It can be performed until the $\mathscr{SF}$ measure is large enough to show that the appropriate discrepancy among the resultant classes has been obtained. Through these aforementioned properties, this objective can be achieved by minimizing the total within-class variance $\nu_{WC}(\boldsymbol{T})$. This can be achieved by the scheme that selects the class with the maximal contribution ($w_k \sigma_k^2$) to the total within-class variance for performing the bi-class partition procedure in each recursion. Thus, the $\mathscr{SF}$ measure will most rapidly reach the maximal increment to satisfy sufficient separability among the resultant classes of pixels. As a result, objects with homogeneous gray intensities will be well-separated.

The class having the maximal contribution of within-class variance $w_k \sigma_k^2$ is denoted by C_p, and it comprises a subset interval of gray intensities represented by C_p : $\{t_p + 1,\ t_p + 2, \ldots,\ t_{p+1}\}$. Then a simple effective *bi-class partition procedure*, as described in Ref. [31], is performed on each determined C_p in each recursion until the separability among all classes becomes satisfactory, i.e. the condition where the $\mathscr{SF}$ measure approximates a sufficiently large value. The class C_p will be divided into two classes C_{p0} and C_{p1} by applying the optimal threshold t_S^* determined by the localized histogram based selection procedure as described in Ref. [31]. The resultant classes C_{p0} and C_{p1} comprise the subsets of gray intensities derived from C_p and can be represented as: C_{p0} : $\{t_p + 1,\ t_p + 2, \ldots,\ t_S^*\}$ and C_{p1} : $\{t_S^* + 1,\ t_S^* + 2, \ldots,\ t_{p+1}\}$. The threshold values determined by this recursive selection strategy is ensured to achieve maximum separation on the resultant segmented classes of gray intensities, and hence satisfactory segmentation results of objects can be accomplished by means of the smallest amount of thresholding levels.

Furthermore, if a region $\Re^{i,j}$ is comprised of a set of pixels with homogeneous gray intensities, most of them are parts of a large homogeneous background region, and thus it is unnecessary to be partitioned to avoid the redundant segmentation for saving the computation costs. For example, Fig. 4(b) is the block region with such characteristics. Therefore, before performing the first partition procedure on the region $\Re^{i,j}$, an investigation of the homogeneity of $\Re^{i,j}$ should be conducted in advance to avoid such redundant segmentation. This condition can be determined by evaluating the following two statistical features: (1) the bi-class $\mathscr{SF}$ measure, denoted as $\mathscr{SF}_b$, which is the $\mathscr{SF}$ value obtained by performing the initial *bi-class partition procedure* on the region $\Re^{i,j}$, i.e. the $\mathscr{SF}$ value associated with the determined threshold t_S^* and (2) the standard deviation, $\sigma_{\Re}$, of the gray intensities of the pixels in the entire region $\Re^{i,j}$. According to the aforementioned properties, the $\mathscr{SF}_b$ value reflects the separability of the statistical distribution of gray intensities of pixels in the entire region $\Re^{i,j}$, and the lower the $\mathscr{SF}_b$ value is, the more indistinct or uniform the distribution is. The standard deviation $\sigma_{\Re}$ represents whether the distribution of gray intensities in $\Re^{i,j}$ is widely dispersed or narrowly aggregated. Therefore, a region $\Re^{i,j}$ is determined to be a *homogeneous* region that comprises a set of homogeneous pixels of a uniform object or parts thereof if both the $\mathscr{SF}_b$ and $\sigma_{\Re}$ features reveal low values. On the other hand, if $\mathscr{SF}_b$ is small but $\sigma_{\Re}$ is large, the region $\Re^{i,j}$ may consist of many indistinct object regions with low separability, and should still undergo a recursive partition process to separate all objects. Based on the above-mentioned phenomenon, a region $\Re^{i,j}$ can be recognized as a *homogeneous* region if the following *homogeneity condition* is satisfied:

$$\mathscr{SF}_b \leqslant \tau_{h0}, \quad \text{and} \quad \sigma_{\Re} \leqslant \tau_{h1} \tag{6}$$

where τ_{h0} and τ_{h1} are pre-defined thresholds. If a region $\Re^{i,j}$ is recognized as a *homogeneous* region, then it does not need to undergo the partition process and hence keeps its pixels of homogeneous objects unchanged to be processed by the next stage.

3.2. Recursive partition process of localized regions

Based on the above-mentioned concepts, the localized automatic multilevel thresholding process is performed by the following recursive steps:

Step 1: To begin, the illumination image **Y** with size $W_{img} \times H_{img}$ is divided into localized block regions $\Re^{i,j}$ with the given size $M_H \times M_V$, as shown in Fig. 3. Here (i,j) are the location indices, and $i = 0, \ldots, N_H$ and $j = 0, \ldots, N_V$, where $N_H = (\lceil W_{img} \,/\, M_H \rceil - 1)$ and $N_V = (\lceil H_{img} \,/\, M_V \rceil - 1)$, which represent the numbers of divided block regions per row and per column, respectively.

Step 2: For each block region $\Re^{i,j}$, compute the histogram of pixels in $\Re^{i,j}$, and then determine its associated standard deviation—$\sigma_{\Re}^{i,j}$ and the bi-class separability measure $\mathscr{SF}_b$; initially, there is only one class $C_0^{i,j}$; let q represent the present amount of classes, and thus set $q = 1$. If the homogeneity condition, i.e. Eq. (6), is satisfied, then skip the localized thresholding process for this region $\Re^{i,j}$ and go to step 7; else perform the following steps.

Step 3: Currently, q classes exist, having been decomposed from $\Re^{i,j}$. Compute the class probability $w_k^{i,j}$, the class mean $\mu_k^{i,j}$, and the standard deviation $\sigma_k^{i,j}$, of each existing class $C_k^{i,j}$ of gray intensities decomposed from $\Re^{i,j}$, where k denotes the index of the present classes and $k = 0, \ldots, q-1$.

Step 4: From all classes $C_k^{i,j}$, determine the class $C_p^{i,j}$ which has the maximal contribution ($w_k^{i,j} \sigma_k^{i,j2}$) of the total within-class variance

$v_{WC}^{i,j}$ of $\Re^{i,j}$, to be partitioned in the next step in order to achieve the maximal increment of $\mathscr{SF}$.

Step 5: Partition $C_p^{i,j}$: $\{t_p^{i,j}+1, t_p^{i,j}+2, \ldots, t_{p+1}^{i,j}\}$ into two classes $C_{p0}^{i,j}$: $\{t_p^{i,j}+1, t_p^{i,j}+2, \ldots, t_S^{i,j*}\}$ and $C_{p1}^{i,j}$: $\{t_S^{i,j*}+1, t_S^{i,j*}+2, \ldots, t_{p+1}^{i,j}\}$, using the optimal threshold $t_S^{i,j*}$ determined by the bi-class partition procedure. Consequently, the gray intensities of the region $\Re^{i,j}$ are partitioned into q+1 classes, $C_0^{i,j}, \ldots, C_{p0}^{i,j}, C_{p1}^{i,j}, \ldots, C_{q-1}^{i,j}$ and then let $q = q$+1 update the record of the current class amount.

Step 6: Compute the $\mathscr{SF}$ value of all currently obtained classes using Eq. (5), if the objective condition, $\mathscr{SF} \geqslant \tau_{SF}$, is satisfied, then perform the following Step 7; otherwise, go back to Step 3 to conduct further partition process on the obtained classes.

Step 7: Classify the pixels of the block region $\Re^{i,j}$ into separate *SR*s, $SR^{i,j,0}$, $SR^{i,j,1}, \ldots,$ $SR^{i,j,q-1}$ corresponding to the partitioned classes of gray intensities, $C_0^{i,j}$, $C_1^{i,j}, \ldots C_{q-1}^{i,j}$, respectively, where the notation $SR^{i,j,k}$ represents the k-th *SR* decomposed from the region $\Re^{i,j}$. Consequently, we obtain

$$\bigcup_{k=0}^{q-1} SR^{i,j,k} = \Re^{i,j} \quad \text{and} \quad SR^{i,j,k_1} \bigcap_{k_1 \neq k_2} SR^{i,j,k_2} = \phi$$

Then, finish the localized thresholding process on $\Re^{i,j}$ and go back to Step 2 and repeat Steps 2–6 to recursively partition the remaining block regions; if all block regions have been processed, go to Step 8.

Step 8: Terminate the segmentation process and deliver all obtained *SR*s of the corresponding block regions.

Here the separability measure threshold τ_{SF} is a pre-defined threshold to determine whether the segmented objects in the block regions are sufficiently separated to satisfy the objective condition. From our experimental analysis on the block regions containing textual objects in the test images, most of them achieve satisfactory segmentation results of homogeneous objects when their resultant $\mathscr{SF}$ values exceed 0.92 after performing the segmentation procedure, and some other complemental experimental analysis described in Ref. [31] also shows similar consequences. Therefore, the value of τ_{SF} is determined as 0.92 to yield satisfactory segmentation results on the block regions. As for the thresholds τ_{h0} and τ_{h1} utilized in the homogeneity condition, we can also determine the suitable values of them by the similar way. By observing the non-textual background regions containing pixels in homogeneous gray intensities, their associated $\mathscr{SF}_b$ features mostly reflect small values of below 0.6, and are also accompanied with the corresponding $\sigma_{\Re}$ standard deviation features that are below 11. Therefore, the values of the thresholds τ_{h0} and τ_{h1} are chosen as 0.6 and 11, respectively, to appropriately detect non-textual homogeneous block regions before performing the thresholding process, and thus some unnecessary segmentations that produce redundant *SR*s can be efficiently avoided for saving computation costs of the localized multilevel thresholding and the following multi-plane region matching and assembling process.

With regard to the size parameters $M_H \times M_V$ of each block region, in order for the localized thresholding process to be more adaptive on the steep gradation situation, and to extract the foreground objects in greater detail, smaller sized block regions are desirable. In this way the small objects can be more clearly segmented, but at the cost of greater computation so as to yield the final results when performing the subsequent multi-plane region matching and assembling process. Therefore, suitable larger values of M_H and M_V should be chosen to moderately localize and accommodate the features of the allowable character size, and so that the contained textual objects in the images can be clearly segmented. Therefore, given an input document image, M_H and M_V should also be automatically determined with respect to its scanning resolution *RES* (pixels per inch) by applying a size mapping parameter τ_d, and can be obtained by

$$M_H = M_V = \tau_d \cdot RES \tag{7}$$

Based on the analysis of typical characteristics of character sizes as described in Ref. [32] and the practice that typical resolutions for scanning most real-life document images may range from 200 to 600 dpi, the value of τ_d is reasonably determined as 0.4 according to the typical allowable character sizes with respect to the scanning resolutions *RES*. In this way, the size of each block region is determined as about $10 \times 10\,\text{mm}^2$ in different scanning resolutions, such as $M_H = M_V = 80$, $M_H = M_V = 120$, and $M_H = M_V = 240$ in 200, 300, and 600 dpi scanning resolutions, respectively. These parameters are determined by conducting experiments involving numerous real-life document samples with various characteristics in our experimental set, so that nearly all foreground and textual objects in various document images can be appropriately separated in the preliminary experiments.

We utilize Fig. 4 as an example of performing the localized automatic multilevel thresholding procedure on several block regions. Here Fig. 4(a) is part of the sectored sample image in Fig. 3. Figs. 4(b), (d), (g), and (l), show the four adjacent block regions, $\Re^{i_1,j_1}, \Re^{i_2,j_1}, \Re^{i_1,j_2}$, and $\Re^{i_2,j_2}$, their corresponding $\mathscr{SF}_b$ and $\sigma_{\Re}$ values, for illustrating the localized thresholding procedure. Here Fig. 4(b) is a homogenous block region, and is properly detected by the homogenous conditions, and therefore its pixels are kept intact in Fig. 4(c). Figs. 4(d), (g), and (l), are the block regions comprised of multiple homogeneous objects. After the localized histogram multilevel thresholding procedure has been performed, different objects in these localized regions are distinctly segmented into separate *SR*s from darkest to lightest, and their corresponding resultant $\mathscr{SF}$ values also approach to be close to 1.0, as shown in Figs. 4(e), (f), (h)–(k), and (m)–(p), respectively.

4. Multi-plane region matching and assembling process

Having decomposed all localized block regions into several separate classes of pixels by the localized multilevel thresholding procedure, various objects embedded or superimposed in different background objects and textures are, respectively, separated into relevant *SR*s. Then we need a methodology for grouping them into meaningful objects, especially textual objects of interest, for further extraction process. Nowadays, concepts of grouping pixels into meaningful regions are widely applied in region-based image segmentation [33,34]. Nevertheless, contemporary pixel-based image segmentation techniques cannot work well for the purpose of segmenting textual objects in complex document images. More commonly, performing pixel-based region segmentation on textual objects may cause extracted printed characters to be fragmented and falsely connected or occluded by non-text pictorial objects or background textures. Moreover, this way suffers heavy computational costs when applied to real-life document images scanned with 200–600 dpi resolutions.

Therefore, there is a need to develop an effective segmentation approach that will deal with regions instead of pixels, to offer a considerable reduction in computational complexity and provide appropriate preservation to the structural characteristics of extracted textual objects, particularly those of small characters with thin strokes. In this section, we present a multi-plane region matching and assembling method, which adopts both the localized spatial dissimilarity relation and the global feature information, to perceptually classify and assemble these obtained *SR*s to compose a set of object planes ($\mathscr{P}_q$) of homogeneous features, especially textual regions of interest. This proposed multi-plane region matching and assembling process is conducted by recursively performing the following three phases—the *initial plane selection phase*, the *matching phase*, and the *plane construction phase*, as illustrated in Fig. 1.

4.1. Overview and basic definitions

To facilitate the matching and assembling process of the *SR*s obtained from the previous procedure, several concepts and definitions on statistical and spatial features for the *SR*s are introduced in this subsection. First, given the localized multilevel thresholding process to segment the $N_H \times N_V$ block regions of the original image into r *SR*s, a hypothetical "*Pool*" is adopted for initially collecting these obtained *SR*s and representing that they are still unclassified into any object planes. Then, the concept *4-adjacent* refers to the situation in which each *SR* has four sides that border the top, the bottom, the left or the right boundary of its adjoining *SR*s. The *SR*s which are comprised of objects with homogeneous features are assembled to form an object plane $\mathscr{P}_q$. An object plane $\mathscr{P}_q$ represents a set of matching *SR*s, and for each pair of *SR*s in $\mathscr{P}_q$, there are some finite chains of *SR*s that connect them so that each successive pair of *SR*s is *4-adjacent*.

Furthermore, each *SR* may comprise several connected object regions of pixels decomposed from its associated block region $\Re^{i,j}$. Thus the pixels that belong to the object regions of a certain *SR* are said to be *object pixels* of this *SR*, while other pixels in this *SR* are *non-object pixels*. The set of the object pixels in an *SR* indexed at (i,j,k) is defined as follows:

$$\boldsymbol{OP}(SR^{i,j,k}) = \{g(SR^{i,j,k},x,y) | \text{The pixel at } (x,y) \text{ is an object pixel in } SR^{i,j,k}\}$$

where $g(SR^{i,j,k},x,y)$ is the gray intensity of the pixel at location (x, y) in $SR^{i,j,k}$, and the range of x is within $[0, M_H - 1]$ and y is within $[0, M_V - 1]$. As well as the total number of object pixels in $SR^{i,j,k}$, i.e. the amount of object pixels in $\boldsymbol{OP}(SR^{i,j,k})$, is represented by $N_{op}(SR^{i,j,k})$. Then, a mean feature $\mu(SR^{i,j,k})$ is also accordingly obtained for each of these *SR*s. Here $\mu(SR^{i,j,k})$ is the mean of gray intensities of object pixels comprised by $SR^{i,j,k}$, and is equivalent to $\mu_k^{i,j}$ obtained in the localized multilevel thresholding process.

Accordingly, given the unclassified *SR*s in the *Pool*, the initial plane selection phase is firstly performed on these unclassified *SR*s to determine a representative set of seed *SR*s $\{SR_m^*,\ m = 0 :\ N-1\}$, and then initially setting up N initial object planes $\{\mathscr{P}_m :\ m = 0 : N-1\}$ based on these selected seed *SR*s. Afterward, the matching phase will be subsequently performed on the rest of unclassified *SR*s in the *Pool* and these initial planes, to determine the association and belongingness of these *SR*s with the existent object planes. For the unclassified *SR*s having perceptibly distinct features with currently existing planes, the plane construction phase will then be conducted to create and initialize an appropriate new plane for assembling *SR*s with such features into this new plane to form another homogeneous object region in the subsequent matching phase recursion. After the first pass of multi-plane region matching and assembling process has been performed, the matching phase and the plane construction phase will be recursively performed in turns on the rest of unclassified *SR*s in the *Pool* and emerging planes, until each *SR* has been classified and associated with a particular plane, and the *Pool* is eventually cleared. As a result, the whole illumination image $\mathbf{Y}$ will be segmented into a set of separate object planes $\{\mathscr{P}_q : q = 0 : L-1\}$, each of which consists of homogenous objects with connected and similar features, such as textual regions of interest, non-text objects such as graphics and pictures, and background textures. Consequently, we will obtain,

$$\bigcup_{q=0}^{L-1} \mathscr{P}_q = \mathbf{Y}, \quad \text{with} \quad \mathscr{P}_{q_1} \bigcap_{q_1 \neq q_2} \mathscr{P}_{q_2} = \phi$$

where L is the number of the resultant planes obtained. In the following subsections, we will, respectively, describe the detailed elements of the proposed multi-plane region matching and assembling process.

4.2. Initial plane selection phase

In this initial processing phase, determining the number and approximate location of the significant clusters of *SR*s in the *Pool* can facilitate the speed and accurateness of the final convergence of the multi-plane region matching and assembling process. For this purpose, the subtractive, and mountain clustering technique [35,36] is applied to determine the *SR*s with the most prominent and representative gray intensity features from the *Pool* set. As a result, the *SR*s being selected as the seeds by the mountain clustering process will be adopted to establish a set of initial object planes for clustering those *SR*s having homogeneous features with them.

The mountain method is a fast, one-pass algorithm, which utilizes the density of features to determine the most representative feature points as the approximate cluster centers. Here we employ the mean features associated with *SR*s, i.e. $\mu(SR)$, as the feature points employed in the mountain clustering process. To facilitate the description of the mountain clustering process, the *region dissimilarity measure*, denoted by D_{RM}, between each pair of the two *SR*s, $SR^{i,j,k}$ and $SR^{i',j',k'}$, is defined as

$$D_{RM}(SR^{i,j,k}, SR^{i',j',k'}) = ||\mu(SR^{i,j,k}) - \mu(SR^{i',j',k'})|| \tag{8}$$

The range of the D_{RM} is within $[0, 255]$. The lower the computed value of D_{RM}, the stronger the similarity among two *SR*s. Then, the initial mountain function at an *SR* is computed as

$$M(SR^{i,j,k}) = \sum_{\forall SR^{i,j,k} \in Pool} e^{-\alpha \cdot D_{RM}(SR^{i,j,k}, SR^{i',j',k'})} \tag{9}$$

where α is a positive constant. It is obvious from Eq. (9) that an *SR* that can attract more *SR*s having similar features to it will obtain a high value in the mountain function. The mountain can be viewed as a measure of the density of *SR*s in the vicinity of the gray intensity feature space. Therefore, it is reasonable to choose *SR*s with the most significant mountain values as representative seeds to create an object plane. Let M_m^* denote the maximal value of the m-th mountain function, and SR_m^* denote the *SR* whose mountain value is M_m^*. They are determined by

$$M_m^* = M_m(SR_m^*) = \max_{\forall SR^{i,j,k} \in Pool} [M_m(SR^{i',j',k'})] \tag{10}$$

First, by applying Eqs. (9) and (10) on all the *SR*s in the *Pool*, we can obtain the first (and highest) mountain M_0^*, and its associated representative *SR*, SR_0^*. Then SR_0^* will be selected as the first seed of the first initial plane. After performing the first iteration of mountain clustering, the following representative seeded *SR*s can be accordingly determined by, respectively, destructing the mountains. This is because the *SR*s whose gray intensity features are close to previously determined seeded *SR*s have influential effects on the values of the subsequent mountain functions, and thus it is necessary to eliminate these effects of the identified seeded *SR*s before determining the follow-up seeded *SR*s. Toward this purpose, the updating equation of the mountain function, after eliminating the last $(m-1)$-th seeded $SR{-}SR_{m-1}^*$, is computed by

$$M_m(SR^{i',j',k'}) = M_{m-1}(SR^{i',j',k'}) - M_{m-1}^* e^{-\beta \cdot D_{RM}(SR^{i,j,k}, SR_{m-1}^*)} \tag{11}$$

where the parameter β determines the neighborhood radius that provides measurable reductions in the updated mountain function.

Accordingly, through recursively performing the discount process of the mountain function given by Eq. (11), new suitable seeded *SR*s can be determined in the same manner, until the level of the current maximal M^*_{m-1} falls bellow a certain level compared to that of the first maximal mountain M^*_0. The terminative criterion of this procedure is defined as

$$(M^*_{m-1}/M^*_0) < \delta \tag{12}$$

where δ is a positive constant less 1. Here the parameters are selected as $\alpha = 5.4$, $\beta = 1.5$ and $\delta = 0.45$ as suggested by Pal and Chakraborty [37]. Consequently, this process converges to the determination of resultant N seeded *SR*s: $\{SR^*_m,\ m = 0 : N-1\}$, and they are utilized to establish N initial object planes $\{\mathscr{P}_m :\ m = 0 : N-1\}$ for performing the following *matching phase*.

4.3. Matching phase

Having a set of existent object planes from the initial processing phase or previous iterations of the assembling process, then an efficient methodology to associate and assemble the unclassified *SR*s remained in the *Pool* with these object planes is necessary to produce appropriate segmentation results of textual objects. Toward this goal, we present a matching process for these unclassified *SR*s to, respectively, evaluate their mutual connectedness and similarity associated with the already existing planes, and to determine its best belonging plane.

4.3.1. Matching grades

To effectively determine the best belonging plane of an unclassified *SR*, we employ a hybrid methodology, named the *matching grade* evaluation, for evaluating the mutual connectedness and similarity between them. This hybrid evaluation methodology considers both local pair-wise and global information provide by *SR*s and existing planes based on two forms of matching grades, the *single-link matching grade*, and the *centroid-link matching grade*. The single-link matching grade is performed by examining the degree of local disconnectedness between a pair of two neighboring *SR*s, an unclassified *SR* and its neighboring classified *SR*s already have belonging planes; while the centroid-link matching grade is adopted for assessing the degree of global dissimilarity between an unclassified *SR* and an already existing plane. Then the two matching grades are combined to provide an effective hybrid criterion to determine the best belonging plane for this unclassified *SR* among all the existing planes.

During one given matching phase recursion, if an unclassified *SR* can find its best belonging plane after examining their mutual matching grade, then this *SR* is classified and assembled into this best belonging plane and removed from the *Pool* afterward; otherwise, if there is no suitable matching plane for an unclassified *SR* at this time, then this *SR* will remain unclassified in the *Pool*. Since new potential object planes will be created in the following recursion of the plane constructing phase, *SR*s remaining unclassified in the current matching phase recursion will be re-analyzed in subsequent recursions until their best matching planes are determined.

The single-link matching grade is utilized to examine the degree of disconnectedness between an unclassified *SR* in the *Pool*, $SR^{i,j,k}$, and an already existent plane $\mathscr{P}_q$ in a local manner. It is determined by applying a connectedness measure on $SR^{i,j,k}$ and its 4-adjacent *SR*s which have already belonged to an existent plane $\mathscr{P}_q$, denoted by $SR_q^{i',j',k'}$, where the subscript q represents that $SR_q^{i',j',k'}$ belongs to the q-th plane $\mathscr{P}_q$. To effectively evaluate the single-link matching grade, two measures for evaluating discontinuity and dissimilarity between a pair of two 4-adjacent *SR*s—the *side-match measure* and the region dissimilarity measure, i.e. D_{RM} as computed using Eq. (8), are employed. Then both D_{SM} and D_{RM} measures are jointly considered to determine the single-link matching grade of a pair of 4-adjacent *SR*s.

The side-match measure, denoted by D_{SM}, which examines the degree of disconnectedness of the touching boundary between $SR^{i,j,k}$ and $SR_q^{i',j',k'}$, is described as follows. Given such pair of two *SR*s are 4-adjacent, they may have one of the two types of touching boundaries: (1) a vertical touching boundary mutually shared by two horizontally adjacent *SR*s, as shown in Fig. 5(a) or (2) a horizontal boundary shared by two vertically adjacent *SR*s, as shown in Fig. 5(b).

First, given a pair of two horizontally adjacent *SR*s—$SR^{i,j,k}$ on the left and $SR_q^{i',j',k'}$ on the right, the gray intensities of pixels on the rightmost side of $SR^{i,j,k}$ and the leftmost side of $SR_q^{i',j',k'}$ can be described as: $g(SR^{i,j,k}, M_H - 1, y)$ and $g(SR_q^{i',j',k'}, 0, y)$, respectively. Then the sets of object pixels on the rightmost side and the leftmost side of a given *SR*, denoted by $\boldsymbol{RS}(SR^{i,j,k})$ and $\boldsymbol{LS}(SR^{i,j,k})$, respectively, are defined as follows:

$$\begin{aligned}\boldsymbol{RS}(SR^{i,j,k}) &= \{g(SR^{i,j,k}, M_H - 1, y) | g(SR^{i,j,k}, M_H - 1, y) \\ &\in \boldsymbol{OP}(SR^{i,j,k}),\ \text{and}\ 0 \leqslant y \leqslant M_V - 1\} \quad \text{and} \\ \boldsymbol{LS}(SR^{i,j,k}) &= \{g(SR^{i,j,k}, 0, y) | g(SR^{i,j,k}, 0, y) \in \boldsymbol{OP}(SR^{i,j,k}), \\ &\text{and}\ 0 \leqslant y \leqslant M_V - 1\}\end{aligned}$$

To facilitate the following descriptions of the side-match features, the denotations of $SR^{i,j,k}$ and $SR_q^{i',j',k'}$ are simplified as SR^l and SR^r, respectively. The vertical touching boundary of SR^l and SR^r, denoted as $\boldsymbol{VB}(SR^l, SR^r)$, is represented by a set of side connections formed by pairs of object pixels that are symmetrically connected on their associated rightmost and leftmost sides, and is defined as follows:

$$\begin{aligned}\boldsymbol{VB}(SR^l, SR^r) &= \{(g(SR^l, M_H - 1, y), g(SR^r, 0, y)) | g(SR^l, M_H - 1, y) \\ &\in \boldsymbol{RS}(SR^l),\ \text{and}\ g(SR^r, 0, y) \in \boldsymbol{LS}(SR^r)\}\end{aligned}$$

Similarly, in the case that $SR^{i,j,k}$ and $SR_q^{i',j',k'}$ are vertically adjacent (suppose that $SR^{i,j,k}$ is on the top and $SR_q^{i',j',k'}$ is on the bottom, and their denotations are also simplified as SR^t and SR^b, respectively), their horizontal touching boundary can be represented as

$$\begin{aligned}\boldsymbol{HB}(SR^t, SR^b) &= \{(g(SR^t, x, M_v - 1), g(SR^b, x, 0)) | g(SR^t, x, M_v - 1) \\ &\in \boldsymbol{BS}(SR^t),\ \text{and}\ g(SR^b, x, 0) \in \boldsymbol{TS}(SR^b)\}\end{aligned}$$

where $\boldsymbol{BS}(SR^t)$ and $\boldsymbol{TS}(SR^b)$ represent the bottommost side and the topmost side of SR^t and SR^b, respectively, and are defined as

$$\begin{aligned}\boldsymbol{BS}(SR^{i,j,k}) &= \{g(SR^{i,j,k}, x, M_v - 1) | g(SR^{i,j,k}, x, M_v - 1) \\ &\in \boldsymbol{OP}(SR^{i,j,k}),\ \text{and}\ 0 \leqslant x \leqslant M_H - 1\} \quad \text{and} \\ \boldsymbol{TS}(SR^{i,j,k}) &= \{g(SR^{i,j,k}, x, 0) | g(SR^{i,j,k}, x, 0) \\ &\in \boldsymbol{OP}(SR^{i,j,k}),\ \text{and}\ 0 \leqslant x \leqslant M_H - 1\}\end{aligned}$$

Also, the number of side connections of the touching boundary, i.e. the amount of connected pixel pairs in $\boldsymbol{VB}(SR^{i_1,j_1,k_1}, SR^{i_2,j_2,k_2})$ or $\boldsymbol{HB}(SR^{i_1,j_1,k_1}, SR^{i_2,j_2,k_2})$, should also be considered for normalizing the disconnectedness measure of the two 4-adjacent *SR*s, and is denoted by $N_{sc}(SR^{i_1,j_1,k_1}, SR^{i_2,j_2,k_2})$.

Therefore, the horizontal and vertical types of the side-match measures of a pair of two 4-adjacent *SR*s, denoted by D^h_{SM} and D^v_{SM}, respectively, can be computed as

$$D^h_{SM}(SR^l, SR^r) = \frac{\sum_{(g(SR^l,M_H-1,y),g(SR^r,0,y))\in \boldsymbol{VB}(SR^l,SR^r)} ||g(SR^l, M_H-1, y) - g(SR^r, 0, y)||}{N_{sc}(SR^l, SR^r)} \quad \text{and}$$

$$D^v_{SM}(SR^t, SR^b) = \frac{\sum_{(g(SR^t,x,M_v-1),g(SR^b,x,0))\in \boldsymbol{HB}(SR^t,SR^b)} ||g(SR^t, x, M_v-1) - g(SR^b, x, 0)||}{N_{sc}(SR^t, SR^b)} \tag{13}$$

Accordingly, the side-match measure of $SR^{i,j,k}$ and $SR^{i',j',k'}_q$ can be obtained by

$$D_{SM}(SR^{i,j,k}, SR^{i',j',k'}_q) = \begin{cases} D^h_{SM}(SR^l, SR^r) & SR^{i,j,k} \text{ and } SR^{i',j',k'}_q \text{ are horizontally adjacent} \\ D^v_{SM}(SR^t, SR^b) & SR^{i,j,k} \text{ and } SR^{i',j',k'}_q \text{ are vertically adjacent} \end{cases} \tag{14}$$

The range of D_{SM} values is within $[0, 255]$. If the D_{SM} value of two 4-adjacent *SR*s is sufficiently low, then these two *SR*s are homogeneous with each other, and thus they should belong to the same plane.

Accordingly, the D_{SM} measure can reflect the disconnectedness of two 4-adjacent *SR*s, and the D_{RM} value, as obtained by Eq. (8), and assesses the dissimilarity between them. The single-link matching grade, denoted by $\mathfrak{m}_s$, evaluates both the degree of disconnectedness and dissimilarity of the two 4-adjacent *SR*s by considering the dominant effect of their associated D_{SM} and D_{RM} values, and is determined by

$$\mathfrak{m}_s(SR^{i,j,k}, SR^{i',j',k'}_q) = \frac{\max(D_{SM}(SR^{i,j,k}, SR^{i',j',k'}_q),\ D_{RM}(SR^{i,j,k}, SR^{i',j',k'}_q))}{\max(\sigma(SR^{i,j,k}) + \sigma(SR^{i',j',k'}_q),\ 1)} \tag{15}$$

where $\sigma(SR^{i,j,k})$ is the standard deviation of gray intensities of all object pixels associated with $SR^{i,j,k}$, and is equivalent to σ^{ij}_k obtained in the localized histogram multilevel thresholding process. Here the denominator term $\max(\sigma(SR^{i,j,k}) + \sigma(SR^{i',j',k'}_q),\ 1)$ in Eq. (15) serves as the normalization factor.

Next, the centroid-link matching grade, which evaluates the degree of dissimilarity between $SR^{i,j,k}$ and an already existing plane $\mathscr{P}_q$ in a global manner, is given as follows. Let $\mu(\mathscr{P}_q)$ and $\sigma^2(\mathscr{P}_q)$ denote the mean and variance of the existing plane $\mathscr{P}_q$, respectively, and they are given by

$$\mu(\mathscr{P}_q) = \frac{\sum_{SR^{i',j',k'}_q \in \mathscr{P}_q} N_{op}(SR^{i',j',k'}_q) \cdot \mu(SR^{i',j',k'}_q)}{N_{op}(\mathscr{P}_q)} \tag{16}$$

and

$$\sigma^2(\mathscr{P}_q) = \frac{\sum_{SR^{i',j',k'}_q \in \mathscr{P}_q} N_{op}(SR^{i',j',k'}_q) \cdot ||\mu(SR^{i',j',k'}_q) - \mu(\mathscr{P}_q)||^2}{N_{op}(\mathscr{P}_q)} \tag{17}$$

where $N_{op}(\mathscr{P}_q)$ denotes the amount of pixels in $\mathscr{P}_q$, and is given by

$$N_{op}(\mathscr{P}_q) = \sum_{SR^{i',j',k'}_q \in_q} N_{op}(SR^{i',j',k'}_q) \tag{18}$$

Accordingly, the centroid-link matching grade of $SR^{i,j,k}$ and $\mathscr{P}_q$ can be computed by

$$\mathfrak{m}_c(SR^{i,j,k}, \mathscr{P}_q) = \frac{||\mu(SR^{i,j,k}) - \mu(\mathscr{P}_q)||}{\max(\sigma(SR^{i,j,k}) + \sigma(\mathscr{P}_q),\ 1)} \tag{19}$$

If $SR^{i,j,k}$ is finally determined to be merged into the plane $\mathscr{P}_q$, then the mean $\mu(\mathscr{P}_q)$ and variance $\sigma^2(\mathscr{P}_q)$ of $\mathscr{P}_q$ should be updated after taking in $SR^{i,j,k}$. The new mean and variance of $\mathscr{P}_q$ are, respectively, computed by

$$\mu(\mathscr{P}^{new}_q) = \frac{(N_{op}(\mathscr{P}^{prev}_q) \cdot \mu(\mathscr{P}^{prev}_q) + N_{op}(SR^{i,j,k}) \cdot \mu(SR^{i,j,k}))}{(N_{op}(\mathscr{P}^{prev}_q) + N_{op}(SR^{i,j,k}))} \tag{20}$$

and

$$\sigma^2(\mathscr{P}^{new}_q) = \frac{[N_{op}(\mathscr{P}^{prev}_q) \cdot \sigma^2(\mathscr{P}^{prev}_q) + N_{op}(SR^{i,j,k}) \cdot ||\mu(SR^{i,j,k}) - \mu(\mathscr{P}^{new}_q)||^2 + N_{op}(\mathscr{P}^{prev}_q) \cdot ||\mu(\mathscr{P}^{new}_q) - \mu(\mathscr{P}^{prev}_q)||^2]}{(N_{op}(\mathscr{P}^{prev}_q) + N_{op}(SR^{i,j,k}))} \tag{21}$$

where $\mathscr{P}^{new}_q$ denotes the newly expanded plane $\mathscr{P}_q$, while $\mathscr{P}^{prev}_q$ denotes the previous one; and $\mu(\mathscr{P}^{new}_q)$ and $s^2(\mathscr{P}^{new}_q)$ represent the updated mean and variance of $\mathscr{P}_q$, respectively, while $\mu(\mathscr{P}^{prev}_q)$ and $\sigma^2(\mathscr{P}^{prev}_q)$ represent the previous ones.

Both of the above-mentioned matching grades are then combined to form a composite matching grade, denoted by $\mathfrak{M}(SR^{i,j,k}, \mathscr{P}_q)$, to complimentarily assess the degree of disconnectedness and dissimilarity of an unclassified *SR* and an already existing plane in both local pair-wise and global manners. Consequently, this composite matching grade can provide a more effective criterion for determining the best belonging plane for each of the unclassified *SR*s. In each recursion of the matching phase, each of the unclassified *SR*s, i.e. $SR^{i,j,k}$ in the *Pool*, is analyzed by evaluating the composite matching grade of $SR^{i,j,k}$ associated with each of its neighboring existent planes $\mathscr{P}_q$, to seek for the best matching plane into which $SR^{i,j,k}$ should belong.

Since the evaluating process of the composite matching grades of $SR^{i,j,k}$ is performed on its neighboring planes, a plane $\mathscr{P}_q$ must have at least one of its own *SR*s 4-adjacent to $SR^{i,j,k}$, to compete for the belongingness of $SR^{i,j,k}$. To facilitate the computation of the composite matching grade of $SR^{i,j,k}$ and a plane $\mathscr{P}_q$, the processing set $\boldsymbol{AS}(SR^{i,j,k}, \mathscr{P}_q)$ is utilized to store the SR_qs which belong to $\mathscr{P}_q$ and 4-adjacent to $SR^{i,j,k}$ as well, and is defined by

$$\boldsymbol{AS}(SR^{i,j,k}, \mathscr{P}_q) = \{SR^{i',j',k'}_q \in \mathscr{P}_q | SR^{i',j',k'}_q \text{ is 4-adjacent to } SR^{i,j,k}\}$$

Then the composite matching grade $\mathfrak{M}$ of $SR^{i,j,k}$ associated with the plane $\mathscr{P}_q$, which reveals how well $SR^{i,j,k}$ matches with $\mathscr{P}_q$, can be determined by

$$\mathfrak{M}(SR^{i,j,k}, \mathscr{P}_q) = w_c(\mathfrak{m}_c(SR^{i,j,k}, \mathscr{P}_q)) + w_s\left(\min_{\forall SR^{i',j',k'}_q \in \boldsymbol{AS}(SR^{i,j,k}, \mathscr{P}_q)} \mathfrak{m}_s(SR^{i,j,k}, SR^{i',j',k'}_q)\right) \tag{22}$$

where w_c and w_s are the weighting factors to control the weighted contributions of the centroid-linkage and single-linkage strengths of the composite matching grade, respectively, and $w_c + w_s = 1$. By applying the weighting factors w_c and w_s in the composite matching grade, the centroid-linkage and single-linkage can be combined for taking advantage of their related strengths. Because textual regions mostly

reveal obvious spatial connectedness, we reasonably strengthen the single-linkage weight of the composite matching grade, and thus the values of the weighting factors are chosen as $w_c = 0.45$ and $w_s = 0.55$, respectively. Besides, if $SR^{i,j,k}$ has no neighboring $SR_q s$ in $\mathscr{P}_q$, i.e. $\boldsymbol{AS}(SR^{i,j,k}, \mathscr{P}_q) = \phi$, then $\mathscr{P}_q$ is excluded from the consideration for matching with $SR^{i,j,k}$, that is, the evaluation process of their composite matching grade is skipped.

4.3.2. Determination of the best belonging plane

As a result, the best candidate belonging plane for $SR^{i,j,k}$, i.e. the plane having the lowest composite matching grade associated with $SR^{i,j,k}$ among all existing planes, denoted by $\mathscr{P}_{\mathfrak{m}}$, can be determined by

$$\mathfrak{M}(SR^{i,j,k}, \mathscr{P}_{\mathfrak{m}}) = \min_{\forall \mathscr{P}_q} \mathfrak{M}(SR^{i,j,k}, \mathscr{P}_q) \tag{23}$$

If the determined value of $\mathfrak{M}(SR^{i,j,k}, \mathscr{P}_{\mathfrak{m}})$ is too large, $SR^{i,j,k}$ is not likely to have sufficient connectedness and similarity to $\mathscr{P}_{\mathfrak{m}}$. The following *matching criterion* is applied to check whether the currently selected candidate plane $\mathscr{P}_{\mathfrak{m}}$ and $SR^{i,j,k}$ are sufficiently matched, and then the suitability of $SR^{i,j,k}$ for belonging to $\mathscr{P}_{\mathfrak{m}}$ can be determined well. This matching criterion is defined as follows:

$$\mathfrak{M}(SR^{i,j,k}, \mathscr{P}_{\mathfrak{m}}) \leqslant \tau_{\mathfrak{m}} \tag{24}$$

where $\tau_{\mathfrak{m}}$ is a predefined threshold which represents the acceptable tolerance of dissimilarity for $SR^{i,j,k}$ to be grouped into $\mathscr{P}_{\mathfrak{m}}$. The matching criterion has a moderate effect on the number of resultant object planes, and the value choice of $\tau_{\mathfrak{m}}=1.5$ is experimentally determined to obtain sufficiently distinct planes and avoid excessive splitting of planes. This selected value of $\tau_{\mathfrak{m}}$ infers that the dissimilarity feature between $SR^{i,j,k}$ and $\mathscr{P}_{\mathfrak{m}}$ should not exceed approximately three times of their average standard deviation, i.e. $(\sigma(SR^{i,j,k}) + \sigma(\mathscr{P}_q))/2$. This choice is motivated by the assumption that all the pixels in $SR^{i,j,k}$ and $\mathscr{P}_{\mathfrak{m}}$ are independent and normally distributed, over 99% of the homogenous gray intensity features are distributed within three standard deviations of the mean. Besides, Chebyshev's theorem assurances that 88.9% of the gray intensity features are within three standard deviations of the mean, regardless of the distribution.

After the above-mentioned determination process, if $SR^{i,j,k}$ and its associated $\mathscr{P}_{\mathfrak{m}}$ can satisfy the matching criterion, then $SR^{i,j,k}$ is merged into $\mathscr{P}_{\mathfrak{m}}$, and removed from the *Pool*. Otherwise, if the matching criterion cannot be satisfied, this reflects that $SR^{i,j,k}$ is distinct from all its existent adjoining planes, and there is no appropriate belonging plane for $SR^{i,j,k}$ during this current matching phase recursion. Therefore, $SR^{i,j,k}$ will remain in the *Pool*, until its suitable matching plane emerges or it begins its own object plane in the following recursions of the plane construction phase. After a belonging determination has been made for $SR^{i,j,k}$, the matching process is in turn applied on the subsequent unclassified *SR*s in the *Pool*, until all the rest unclassified *SR*s have been processed one time in the current matching phase recursion.

4.4. Plane construction phase

After one given matching phase recursion has been performed, if there are unclassified *SR*s remaining and the *Pool* is not drained as well, these unclassified *SR*s need to be analyzed to determine whether it is necessary to establish a new object plane to assemble the *SR*s with such features into this new plane to form another homogeneous object region. Toward this goal, it is reasonable to find an *SR* having sufficient distinctive features and obviously different features compared to any existent planes, to start a new meaningful object plane. Therefore, in the next matching phase recursion, the follow-up unclassified *SR*s having homogeneous features with this newly created plane can be successively assembled. Since textual regions and homogeneous objects may contain several connected regions, the plane construction phase will determine whether to (1) create and initialize a new plane by selecting the unclassified *SR* having "farthest" features from all existing planes as an initial seed or (2) expand one suitably selected existent plane by merging one unclassified *SR* having "closest" features to this plane, to avoid unnecessary split of homogeneous object regions. The determination is made according to the analysis of the gray intensity and spatial location features described in the following subsection.

4.4.1. Gray intensity and spatial location features of unclassified SRs and existent object planes

The dissimilarity between one unclassified *SR*, which is not adjoined to any one of the currently existing planes in the previous matching phase recursion, and a certain object plane $\mathscr{P}_q$, can be determined by their associated centroid-link matching grade, as computed by Eq. (19). Then the plane having the shortest feature distance to $SR^{i,j,k}$ in gray intensity among all already existing planes, i.e. the plane has the least value of the centroid-link matching grade associated with $SR^{i,j,k}$, denoted by $\mathscr{P}_S(SR^{i,j,k})$, can be obtained by

$$\mathfrak{m}_c(SR^{i,j,k}, \mathscr{P}_S(SR^{i,j,k})) = \min_{\forall \mathscr{P}_q}(\mathfrak{m}_c(SR^{i,j,k}, \mathscr{P}_q)) \tag{25}$$

Here $\mathfrak{m}_c(SR^{i,j,k}, \mathscr{P}_S(SR^{i,j,k}))$ also represents the measure of the least dissimilarity between $SR^{i,j,k}$ and all already existing planes. If $SR^{i,j,k}$ can find a plane $\mathscr{P}_S(SR^{i,j,k})$ having sufficiently low dissimilarity associated with $SR^{i,j,k}$ in gray intensity, and they are also locatively closed as well, then this condition reveals that $SR^{i,j,k}$ is sufficiently homogeneous with $\mathscr{P}_S(SR^{i,j,k})$, even if it is not currently 4-adjacent to $\mathscr{P}_S(SR^{i,j,k})$. This situation should be resolved by detecting and expanding an existent plane having sufficient closing distances in features of gray intensity and spatial location to some unclassified *SR*s, and then this selected plane can take in such unclassified *SR*s in the subsequent matching phase recursion.

For the purpose of dealing with the above-mentioned situation, the locative distance between $SR^{i,j,k}$ and a plane $\mathscr{P}_q$, denoted as $D_E(SR^{i,j,k}, \mathscr{P}_q)$, is computed by the Euclidean distance between $SR^{i,j,k}$ and its closest SR_q among all $SR_q s$ associated with the plane $\mathscr{P}_q$; and is determined as

$$D_E(SR^{i,j,k}, P_q) = \min_{\forall SR_q \in \mathscr{P}_q} D_e(SR^{i,j,k}, SR_q^{i',j',k'}) \tag{26}$$

where

$$D_e(SR^{i,j,k}, SR_q^{i',j,k'}) = \sqrt{(i-i')^2 + (j-j')^2)} \tag{27}$$

If $SR^{i,j,k}$ and its $\mathscr{P}_S(SR^{i,j,k})$ are homogeneous in gray intensity and also locatively close to each other, i.e. both $\mathfrak{m}_c(SR^{i,j,k}, \mathscr{P}_S(SR^{i,j,k}))$ and $D_E(SR^{i,j,k}, \mathscr{P}_S(SR^{i,j,k}))$ values are sufficiently low, then $SR^{i,j,k}$ should join the plane $\mathscr{P}_S(SR^{i,j,k})$, rather than establish a new separate plane, so as to prevent a large textual region or homogeneous object to be split into multiple planes. Otherwise, if no such planes are found, a new plane should be created to aggregate those *SR*s with distinct features.

In order to guarantee that the newly created plane contains distinct features with the current existent planes, a scheme for selecting a suitable *SR* as the representative seed for constructing a new plane is given as follows:

$$\mathfrak{m}_c(SR_{NP}, \mathscr{P}_S(SR_{NP})) = \max_{\forall SR^{i,j,k} \in \boldsymbol{Pool}} \mathfrak{m}_c(SR^{i,j,k}, \mathscr{P}_S(SR^{i,j,k})) \tag{28}$$

In this way, this determined seed *SR*, denoted by SR_{NP}, is the one having most dissimilar features in gray intensity to any already

existing planes, and thus SR_{NP} will begin its own new plane to aggregate those *SR*s whose features are distinct from other existing planes but homogenous with SR_{NP}.

4.4.2. Determination process of new plane construction

By means of the definitions given above, the plane construction phase is performed according to the following steps:

Step 1: First, the unclassified *SR*s having sufficiently low $\mathfrak{m}_c(SR^{i,j,k}, \mathscr{P}_S(SR^{i,j,k}))$ values are selected into the set $\boldsymbol{SR_S}$ through the following operation:

$$\boldsymbol{SR_S} = \{SR^{i,j,k} \in Pool | \mathfrak{m}_c(SR^{i,j,k}, \mathscr{P}_S(SR^{i,j,k})) \leqslant \tau_S\} \tag{29}$$

where τ_S is a predefined threshold for determining whether one *SR* is sufficiently homogeneous with any one of existing planes. If none of the unclassified *SR*s can satisfy the above condition to be selected into $\boldsymbol{SR_S}$, i.e. $\boldsymbol{SR_S} = \phi$, then go directly to Step 3 to construct a new plane; otherwise, perform the following Step 2.

Step 2: The set $\boldsymbol{SR_S}$ now contains the *SR*s being significantly homogeneous with some already existing planes, but are not 4-adjacent with them, and thus remain unclassified in the previous matching phase recursion. The *SR* being locatively nearest to its associated $\mathscr{P}_S(SR^{i,j,k})$, denoted by SR_P, should be determined as

$$D_E(SR_P, \mathscr{P}_S(SR_P)) = \min_{\forall SR^{i,j,k} \in \boldsymbol{SR_S}} D_E(SR^{i,j,k}, \mathscr{P}_S(SR^{i,j,k})) \tag{30}$$

If SR_P and its associated $\mathscr{P}_S(SR^{i,j,k})$ are sufficiently close to each other, i.e. the condition $D_E(SR_P, \mathscr{P}_S(SR_P)) \leqslant \tau_L$ is satisfied, then SR_P is determined to be merged with $\mathscr{P}_S(SR^{i,j,k})$ to enlarge its influential area on nearby *SR*s, and proceeds to Step 4. Otherwise, perform Step 3 to create a new plane.

Step 3: The SR_{NP}, the *SR* most dissimilar to any currently existing planes, is determined by using Eq. (28). Thus, SR_{NP} is employed as a seeded *SR* to create a new plane $\mathscr{P}_{new}$, and then continues to Step 4.

Step 4: Finish the plane construction phase, and then conduct the next matching phase recursion.

The threshold τ_S utilized in Eq. (29) moderately influences the number of resultant planes. If the value of τ_S is low, then some homogeneous textual regions may still be disjoined planes and the number of resultant planes cannot be efficiently reduced, although its influence on text extraction results is insignificant, the increased number of resultant planes will cost more recursions of text extraction process for extracting texts in these planes. Conversely, if the value is too large, although the number of resultant planes and overall computational costs can be reduced, the text extraction performance may be affected because some textual regions may be over-merged with some other non-textual regions. Reasonably, τ_S should be tighter than the value of the matching criterion parameter $\tau_{\mathfrak{m}}$ (in Eq. (24)), to ensure that the determined SR_P is sufficiently homogeneous with its associated $\mathscr{P}_S(SR^{i,j,k})$, and thus the inappropriate expansion of object planes, which may cause some inhomogeneous regions to be absorbed in the subsequent matching phase recursions, can also be avoided. Thus $\mathscr{P}_S(SR^{i,j,k})$ can appropriately attract homogeneous *SR*s near the extended influential area which are benefited from participation with SR_P. Therefore, in our experiments, the value of τ_S is chosen as $\tau_S = 0.8 \cdot \tau_{\mathfrak{m}}$. In general, text-lines or text-blocks usually occupy a perceptible area of the image, and thus their width or height should be in appreciable proportion to those of the whole image. Hence, $\tau_L = \min(N_H, N_V)/4$ is used for experiments, where N_H and N_V are the numbers of block regions per row and per column, respectively.

Consequently, for the unclassified *SR*s having obviously different features with currently existent planes and left in *Pool* after the previous matching phase recursion, the plane construction phase will be conducted to obtain a suitable determination for creating an appropriate new object plane or expand one suitably selected existent plane. Thus, in the subsequent matching phase recursion, the follow-up unclassified *SR*s owning such features associated with this newly created plane can be successively assembled into a whole textual region or a homogeneous object.

4.5. Overall processing

Fig. 6 shows an example of processing results of the sample image in Fig. 2 by the proposed multi-plane segmentation process. In the sample image of Fig. 2, three different colored textual regions were printed on a varying and shaded background, and moreover, the black characters are superimposed on the bright characters. Then, having the *SR*s obtained from the localized multilevel thresholding procedure as illustrated in Fig. 4, the region matching and assembling process is then applied on these obtained *SR*s. This process begins by applying the initial plane selection phase on these *SR*s, and five representative seed *SR*s, $\{SR^*_m,\ m = 0:4\}$, are selected to create the resultant planes $\mathscr{P}_0 - \mathscr{P}_4$, which are the resultant planes shown in Figs. 6(a)–(e). Then the other *SR*s in the *Pool* are analyzed and assembled by recursively performing the matching phase and the plane construction phase, to associate them with the initial planes $\mathscr{P}_0 - \mathscr{P}_4$ and the subsequent emerging planes. As a result, seven major resultant object planes $\mathscr{P}_0 - \mathscr{P}_6$ (while those insignificant planes are discarded) are obtained after completing the multi-plane segmentation process, as depicted in Figs. 6(a)–(g). Within these object planes, the planes $\mathscr{P}_1$, $\mathscr{P}_3$, and $\mathscr{P}_4$ in Figs. 6(b), (d), and (e), respectively, contain textual objects of interest.

Accordingly, the homogeneous objects in which all textual objects and background textures are segmented into several separate object planes can be efficiently analyzed in detail. By observing these obtained planes, we can see that three textual regions with different characteristics are distinctly separated. Thus, extraction of textual objects from each binarized plane in which textual objects are well-separated can be easily performed by some contemporarily developed text extraction techniques. The following section describes a simple procedure for locating and extracting textual objects from these resultant object planes.

5. Text extraction

Having performed the multi-plane segmentation process, the entire image is decomposed into various object planes. Each object plane may consist of various considerable objects, such as textual regions, graphical and pictorial objects, background textures or other objects. Here each individual object plane $\mathscr{P}_q$ will be binarized by setting its object pixels to black, and setting other non-object pixels to white, and hence a "*binarized plane*", denoted as $\mathscr{BP}_q$, is created corresponding to each plane $\mathscr{P}_q$. The text location and extraction process will be performed on each individual binary plane $\mathscr{BP}_q$ to obtain the textual objects of interest. To obtain the character-like components from each binary plane $\mathscr{BP}_q$, a fast connected-component extraction technique [38] is first carried out to locate the connected-components of the black pixels in $\mathscr{BP}_q$. These connected-components may represent the character components, graphical and pictorial objects, or background textures. By extracting the connected-components, the location and dimension of each connected-component are obtained as well. The location and dimension of a connected-component are represented by the bounding box enclosing it.

In this study, based on the concepts of the recursive *XY*-cut techniques for connected-component projections [39,40], we adopt a recursive *XY*-cut spatial clustering process for grouping the connected-components into meaningful sets in each of the binarized planes. We are interested in looking for horizontal text-lines, and hence the

XY-cut spatial clustering process is conducted to cluster the connected-components into several sets in horizontal direction. A resultant set of connected-components may comprise a character string, a larger graphical object, or a group of isolated background components inside the character strings. Each of these connected-component sets, denoted as ***CS***, is then processed by the text identification process to determine whether they are actual text-lines.

Fig. 7 depicts the text extraction process. As shown in Fig. 7(a), for the corresponding connected-components of characters in the binary plane $\mathscr{BP}_4$ (corresponding to the plane $\mathscr{P}_4$ in Fig. 6(e)), there are five resultant ***CS***s obtained after the *X*-cut procedure is performed. The *Y*-cut procedure is in turn performed on these five ***CS***s. For instance, as shown in Fig. 7(b), the *Y*-cut procedure is performed on the ***CS*** at the top of the five ***CS***s obtained from the *X*-cut procedure, and then one resultant ***CS*** is obtained. This is because the connected-components in Fig. 7(b) are all close to each other, and hence are clustered into a single resultant ***CS***. After the *XY*-cut connected-component clustering process on a binary plane is completed, several final ***CS***s are obtained, representing candidates of actual text-lines, as shown in Fig. 7(c). Accordingly, the ***CS***s associated with the remaining binary planes are also obtained after the *XY*-cut process is in turn performed on all binary planes.

The text identification process is then conducted to distinguish whether each one of these obtained ***CS***s comprises actual text-lines or non-text objects. Before distinguishing and extracting text-lines, we first identify halftone pictorial objects and background regions using normalized correlation features [41]. For each one of these ***CS***s, its associated normalized correlation features are computed on the bounding box region covered by its contained components. If the normalized correlation features of one ***CS*** meet the discrimination rules of halftone pictorial objects as suggested in Ref. [41], then it is determined to be a pictorial object or a background region.

After pictorial objects and background regions are identified and eliminated, the text identification is then performed on the rest of ***CS***s. If a ***CS*** actually comprises a text-line, it may have the following distinguishing characteristics: (1) its contained connected-components should be, respectively, aligned, and the number of them should also be in proportion to the width of the ***CS*** and (2) the contained object pixels in the enclosing region of this ***CS*** show distinctive spatial variation. For the first characteristic, the identification strategies of the statistical features of connected-components [24] can be applied to each of the ***CS***s. The second characteristic can be determined by applying the discrimination rules of transition features on the object pixels contained in each of the ***CS***s [42]. Both are independent of font types, lengths and sizes of text strings. Accordingly, a ***CS*** can be identified as an actual text-line if it satisfies both of the above characteristics.

Based on the above-mentioned concepts, the text identification process employs the following determination rules $R_1 - R_5$ for identifying whether one ***CS*** consists of a text-line or non-text objects. A ***CS*** is identified as a real text-line if all of the following decision rules are satisfied. First, the ratio of the width W and the height H of the enclosing bounding box of a ***CS*** should satisfy the size-ratio condition

$$R_1: \ W(CS)/H(CS) \geqslant \tau_r \tag{31}$$

where the threshold τ_r on the size-ratio condition is selected as 2.0 to suitably reflect the rectangular-shaped appearance of a text-line. The number of contained connected-components N_{cc} of a ***CS*** should satisfy the condition,

$$R_2: \ \tau_{n1} \cdot (W(CS)/H(CS)) \leqslant N_{cc}(CS) \leqslant \tau_{n2} \cdot (W(CS)/H(CS)) \text{ and } N_{cc}(CS) > \tau_{n3} \tag{32}$$

where the values of τ_{n1}, τ_{n2} and τ_{n3} are chosen as 0.5, 8.0, and 3, respectively, according to our analysis of typical arrangement and quantity characteristics of the characters in a text-line. Then the ratio of the total area of the bounding boxes of contained connected-components of a ***CS*** to the area of its enclosing box should meet the condition,

$$R_3: \ \tau_{a1} \leqslant \frac{\sum_{C_i \in CS} A(C_i)}{W(CS) \cdot H(CS)} \leqslant \tau_{a2} \tag{33}$$

where C_i is the i-th connected-component contained by the ***CS***, and $A(C_i)$ is the area of the bounding box of C_i; and the thresholds τ_{a1} and τ_{a2} are, respectively, determined as 0.5 and 0.95, to reveal the alignment property of the characters appearing in a text-line. Then the identification rules based on the statistical features of the contained pixels of the ***CS*** are introduced as follows. Considering that "0" represents object pixels and "1" background pixels, the number of transition pixels T_p in the enclosing box of the ***CS*** is determined by computing the number of "0" to "1" and "1" to "0" transitions. Hence the horizontal transition pixel ratio of the ***CS*** must satisfy the condition

$$R_4: \ \tau_{t1} \leqslant T_p(CS)/N_{Col}(CS) \leqslant \tau_{t2} \tag{34}$$

where N_{Col} is the number of the column lines in which the object pixels are present; the values of τ_{t1} and τ_{t2} are chosen as 1.2 and 3.6, respectively, which reflect typical pixel transition features of character strokes. In addition, the density of object pixels in the ***CS*** should also satisfy the condition

$$R_4: \ \tau_{d1} \leqslant \frac{\sum_{C_i \in CS} O_p(C_i)}{W(CS) \cdot H(CS)} \leqslant \tau_{d2} \tag{35}$$

where $O_p(C_i)$ is the number of object pixels of the i-th connected-component of the ***CS***; and the thresholds τ_{d1} and τ_{d2} are set as 0.3 and 0.8, respectively, to reveal the typical occupation characteristic of pixels contained in a text-line.

The above-mentioned discrimination rules and parameters are obtained by analyzing many experimental results of processing document images having text strings with various types, lengths and sizes, and can yield good performance in most general cases. After the text location and identification process has been accordingly applied on all binarized object planes, the text-lines extracted from these planes are then composed into a resultant textual plane, as shown in Fig. 7(d).

6. Experimental results

In this section, the performance of the proposed multi-plane approach is evaluated and compared to several other well-known text extraction techniques, namely Jain and Yu's color-quantization-based method [24], and Pietikainen and Okun's edge-based method [22]. Two test databases of totally 65 real-life complex document images, the first test database consists of 28 English document images, and the second one comprises of 37 Chinese and mixed-Chinese/English document images, are employed for experiments on performance evaluation of text extraction. These test images include a variety of book covers, book, and magazine pages, advertisements, and other real-life documents at the scanning resolution of 200–600 dpi. These images are comprised of textual objects in various colors or illuminations, font styles, and sizes, including sparse and dense textual regions, adjoined or overlapped with pictorial, watermarked, textured, shaded, or uneven illuminated objects and background regions.

6.1. Parameter adaptation

Prior to conduct the text extraction performance evaluation of the proposed approach on the test database, some important parameters utilized in the multi-plane segmentation process should be analyzed and adapted. For this purpose, we will investigate the effect

Table 2
Segmentation accuracy rates with varying τ_{SF} values.

Threshold τ_{SF}	0.84	0.86	0.88	0.9	0.92	0.94	0.96
Accuracy rate (%)	85.1	89.6	92.1	97.9	99.6	97.1	84.6

of these parameters by the experiments of varying one parameter for performance evaluation at a time.

6.1.1. Determination of thresholds for the localized multilevel thresholding

First, to investigate the effect of the threshold settings on the segmentation performance of the localized histogram multilevel thresholding process, a measure that evaluates the segmentation accuracy of the *SR*s obtained from block regions which comprise of textual objects of interest rather than the overall text extraction results. For this purpose, the segmentation accuracy rate is computed by manually counting the number of expected *SR*s comprising textual objects, and the number of adequately segmented *SR*s comprising textual objects in the test images, respectively, and is defined as

$$\text{Accuracy rate} = \frac{\text{No. of correctly segmented textual } SRs}{\text{No. of expected textual } SRs} \tag{36}$$

Since we have already observed that satisfactory segmentation results of homogeneous objects are mostly obtained when their resultant $\mathscr{SF}$ values exceed 0.92 according to the experimental analysis described in our previous study [31], and thus the experiments for evaluating suitable values of the threshold τ_{SF} can be conducted based on this phenomenon. Table 2 depicts the segmentation accuracy rate as τ_{SF} is varied from 0.84 to 0.96 when the size parameters M_H and M_V of block regions are set according to Eq. (7). As can be seen from Table 2, the segmentation accuracy rate is significantly decreased when the value of τ_{SF} is below 0.86 due to growing under-segmented *SR*s, while the accuracy rate is again decayed when τ_{SF} is above 0.94 owing to increasing over-segmented *SR*s. Thus, we can find that the best segmentation performance can be obtained when adopting $\tau_{SF} = 0.92$.

Moreover, the thresholds of the homogeneity condition in Eq. (6), τ_{h0} and τ_{h1}, are adopted for avoiding the redundant segmentation on the non-textual homogeneous block regions because the associated $\mathscr{SF}$ and $\sigma_{\Re}$ values of such block regions are usually significantly lower than those of the block regions containing textual objects. Therefore, an effective selection for the values of τ_{h0} and τ_{h1} should be relatively small and sparse enough not to miss the needed segmentation on textual regions, and yet also moderately large and dense enough to ensure that the unnecessary segmentations on non-textual regions can be sufficiently avoided for saving the computation costs on the localized multilevel thresholding procedure and the following multi-plane region matching and assembling process. This study finds that choosing the values of τ_{h0} and τ_{h1} to be 0.6 and 11, respectively, is efficient for detecting non-textual block regions.

6.1.2. Evaluation of block region size

To examine the effect of the size mapping parameter τ_d and the associated size $M_H \times M_V$ of each block region on the text extraction performance (Eq. (7)), an experiment that evaluating the recall rates of the text extraction results under varying values of τ_d is conducted. The recall rate of the text extraction results is defined as,

$$\text{Recall rate} = \frac{\text{No. of correctly extracted characters}}{\text{No. of actual characters}} \tag{37}$$

Here the recall rate is determined by manually counting the number of total actual characters of the document image, and the number of total correctly extracted characters in the test images, respectively.

Table 3
Text extraction recall rates and average computing times with varying τ_d values.

Size mapping parameter τ_d	0.25	0.3	0.35	0.4	0.45	0.5	0.55
Recall rate (%)	99.1	99.3	99.2	99.2	98.4	97.2	95.7
Average processing time (s)	7.98	3.86	2.09	1.2	0.86	0.55	0.41

Table 4
Text extraction recall rates and average computing times with varying $\tau_{\mathfrak{m}}$ values.

Matching criterion parameter $\tau_{\mathfrak{m}}$	0.9	1.1	1.3	1.5	1.7	1.9	2.1
Recall rate (%)	95.7	98.0	98.9	99.2	98.8	98.3	96.8
Average processing time (s)	3.61	3.39	3.15	2.67	2.5	2.09	1.71

Although the use of small block regions can extract the textual objects in greater detail and provide more adaptiveness on the steep gradation situation during the localized thresholding process to achieve higher text extraction capability of various complex images, if the block region size $M_H \times M_V$ is too small, the increasing number of *SR*s will result in considerably higher computation costs on the localized thresholding process and the multi-plane region matching and assembling process. In contrast, if the block region size is too large, the errors in text extraction results will grow because some detailed information of textual objects may not be clearly segmented during the localized thresholding process. Therefore, determining the appropriate block region size with a suitable value of τ_d is a trade-off between the accuracy of text extraction and the computation costs. According to the aforementioned analysis of typical characteristics of character sizes in the real-life document images, the experiment of evaluating the effect of different block region size is conducted by varying the value of τ_d on the test images scanned by different typical resolution of 200–600 dpi. The results of text extraction recall rates and average computing times as τ_d being varied from 0.25 to 0.55 are depicted in Table 3. As can be seen from Table 3, while a higher recall rate tendency is observed with the smaller block regions, the best selection for both text extraction performance and computation costs can be obtained by using $\tau_d = 0.4$.

6.1.3. Selection of parameters for the multi-plane region matching and assembling process

In this subsection, the appropriate settings of the parameters ($\tau_{\mathfrak{m}}$, τ_S, and τ_L) are analyzed for determining the most appropriate settings for the multi-plane region matching and assembling process. Since the parameters τ_S and τ_L for the plane construction have no significant effect on the text extraction performance, the effect of the matching criterion parameter $\tau_{\mathfrak{m}}$ for the matching phase should be mainly evaluated in this experiment. For this purpose, the influences of the parameters τ_S and τ_L are firstly disabled by setting their values to be 0 and 1, respectively, that is, the unclassified *SR*s having diverse features with any neighboring existent planes will always start new planes in each recursion of the multi-plane region matching and assembling process. Then, the effect of the matching criterion parameter $\tau_{\mathfrak{m}}$ is evaluated by analyzing the recall rates of text extraction results (Eq. (37)) and average computing times under varying values of $\tau_{\mathfrak{m}}$.

Table 4 shows the results of extraction recall rates as $\tau_{\mathfrak{m}}$ is varied from 0.9 to 2.1. From Table 4, we can observe that the text extraction accuracy is gradually decayed when the value of $\tau_{\mathfrak{m}}$ is below the optimal one, as more textual regions are spilt into some superfluous object planes; while the text extraction accuracy is again decreased when the value of $\tau_{\mathfrak{m}}$ is above the optimal one, as more different textual regions are over-grouped into the same object planes.

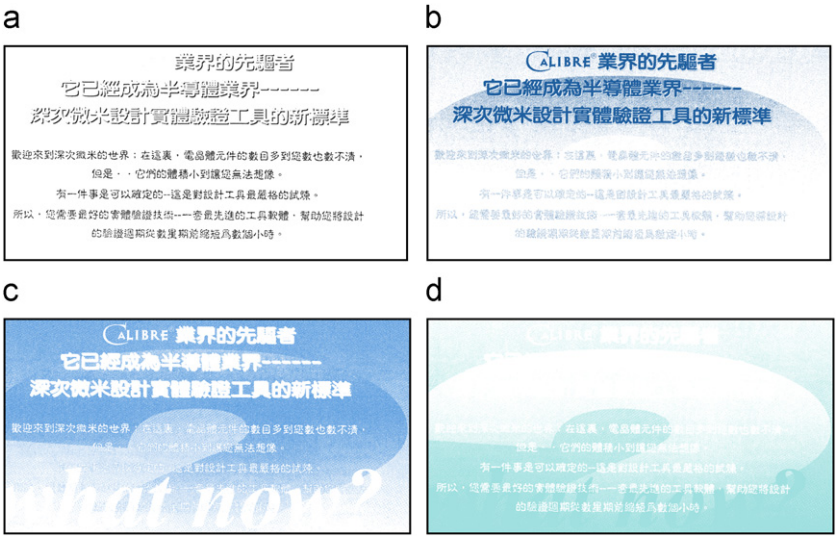

Fig. 8. Representative color images of Fig. 2 after performing Jain and Yu's method: (a) representative color image 1, (b) representative color image 2, (c) representative color image 3, and (d) representative color image 4.

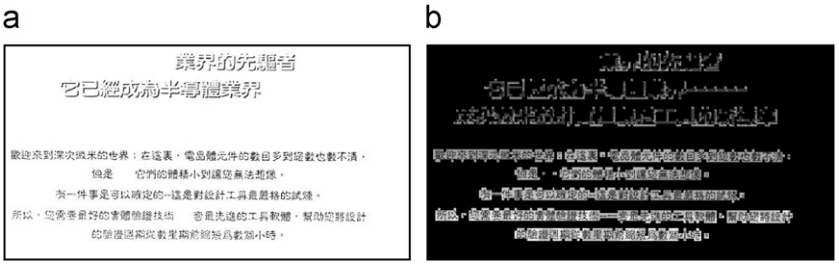

Fig. 9. Text extraction results of Fig. 2 by Jain and Yu's method and Pietikainen and Okun's method: (a) text extraction results by Jain and Yu's method and (b) text extraction results by Pietikainen and Okun's method.

Fig. 10. Original images of the test images 2 and 3: (a) test image 2 (size: 2333×3153) and (b) test image 3 (size: 2405×3207).

Therefore, while the lesser computation time is observed with the larger value of τ_m because of the lesser amount of resultant planes, the best selection for both text extraction performance and computation costs can be obtained with $\tau_m = 1.5$.

The parameters τ_S and τ_L utilized in the plane construction phase are mainly adopted for avoiding the redundant separation of object planes and thus reducing the necessary recursions of text extraction process on obtaining textual objects in these planes for

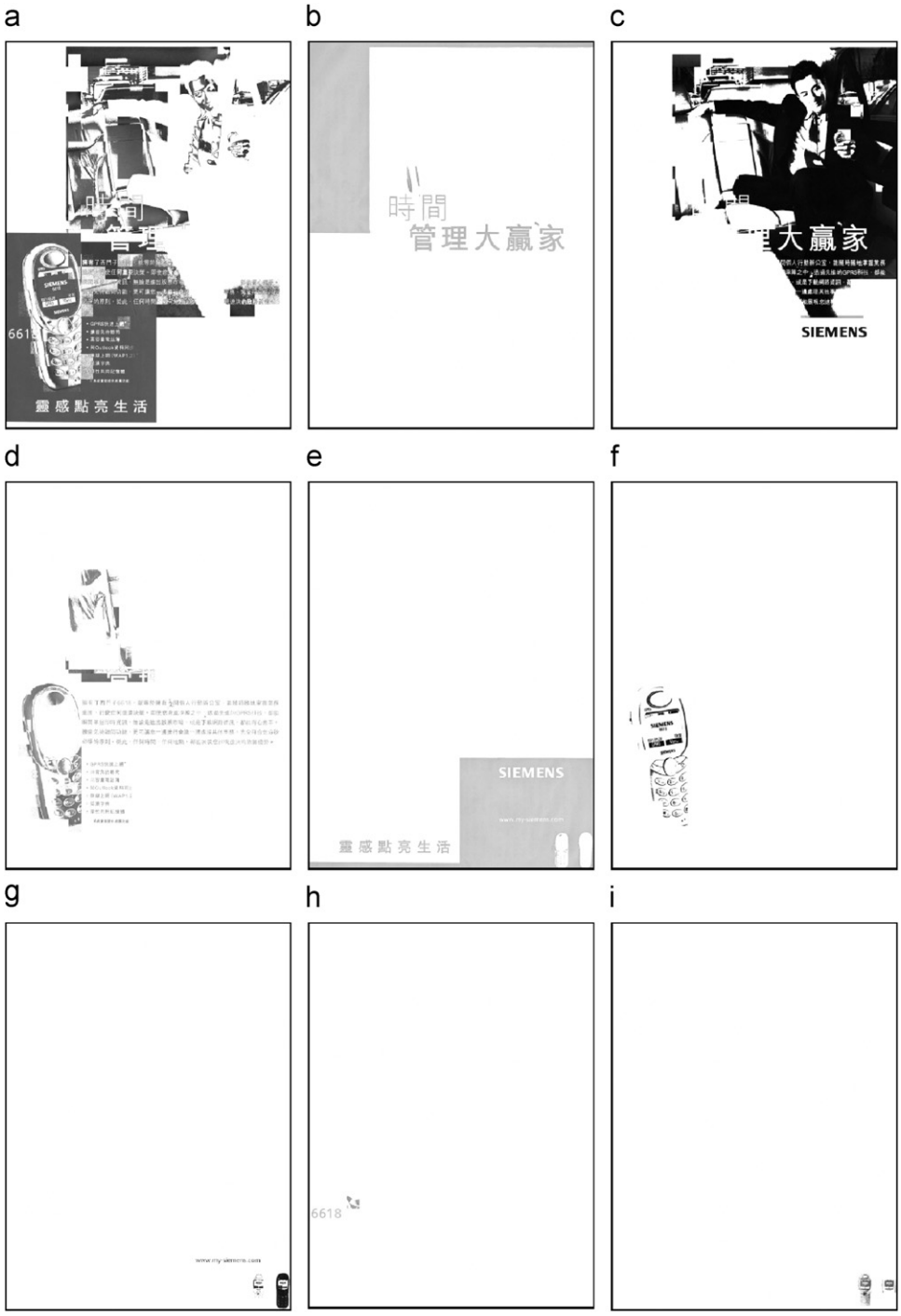

Fig. 11. Decomposed object planes of Fig. 10(a) after performing the proposed multi-plane segmentation: (a) decomposed object plane 1, (b) decomposed object plane 2, (c) decomposed object plane 3, (d) decomposed object plane 4, (e) decomposed object plane 5, (f) decomposed object plane 6, (g) decomposed object plane 7, (h) decomposed object plane 8, and (i) decomposed object plane 9.

saving the computation costs. Variations in the values of the parameters τ_S and τ_L do not significantly affect the text extraction results, except when they are so large that more textual regions being over-merged with some other non-textual regions. Therefore, an effective selection for the values of τ_S and τ_L should be moderately capacious and large enough to keep most homogeneous textual regions in the same planes to ensure fast processing, yet also reasonably tight and small enough to avoid text extraction errors due to textual regions being over-merged with non-textual regions. Based on various experiments with the test images, the current study find that $\tau_S = 0.8 \cdot \tau_m$ with $\tau_L = \min(N_H, N_V)/4$ can sufficiently save the number of object planes for performing text extraction recursion and ensure the satisfactory text extraction accuracy. By using these settings of τ_S and τ_L for the plane construction phase, about 60% of the computation cost for the entire process is saved, that is, the average computation time is reduced from 2.67 s to about 1.2 s when the previously determined optimal parameters are applied.

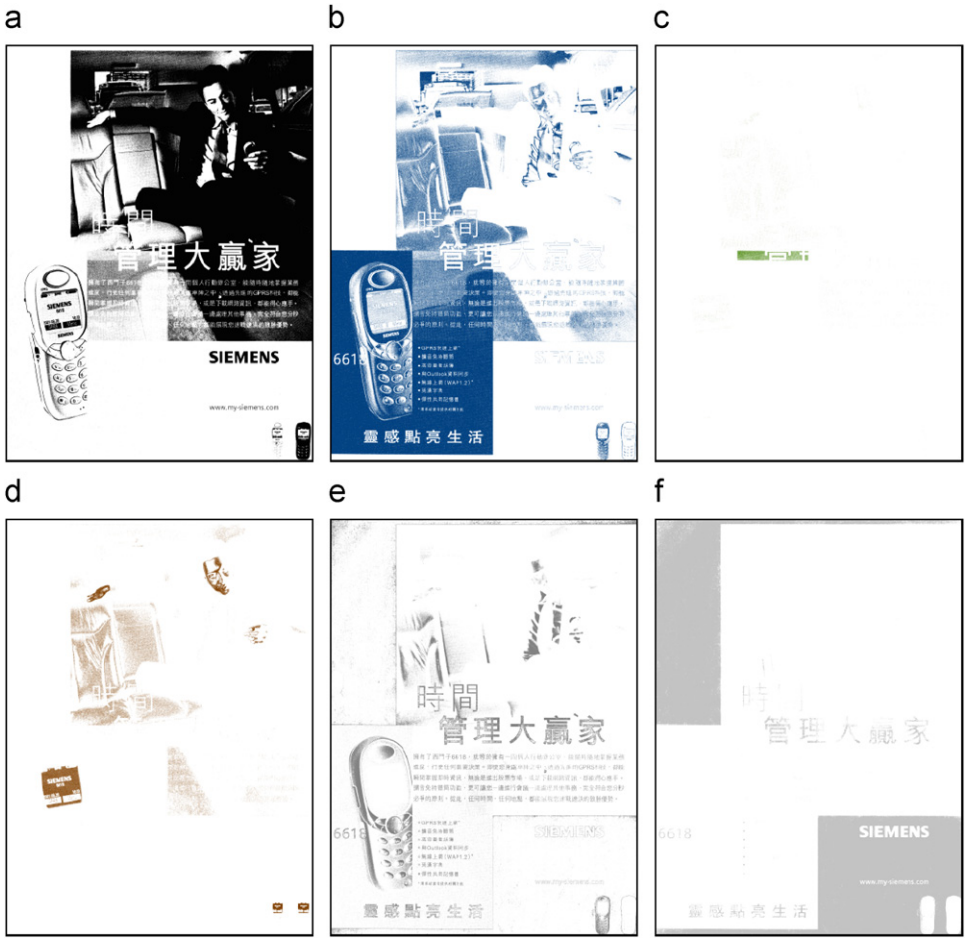

Fig. 12. Representative color images of Fig. 10(a) after performing Jain and Yu's method: (a) representative color image 1, (b) representative color image 2, (c) representative color image 3, (d) representative color image 4, (e) representative color image 5, and (f) representative color image 6.

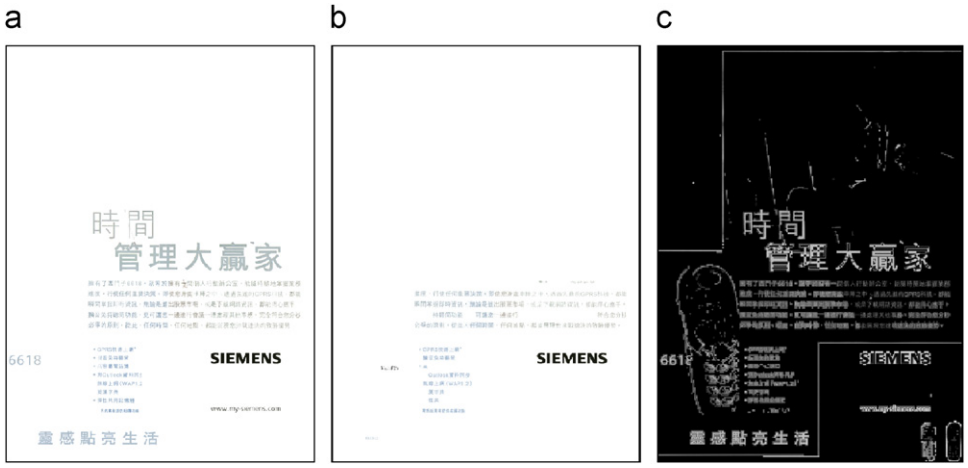

Fig. 13. Text extraction results of Fig. 10(a) by the proposed approach, Jain and Yu's method, and Pietikainen and Okun's method: (a) text extraction results by the proposed approach, (b) text extraction results by Jain and Yu's method, and (c) text extraction results by Pietikainen and Okun's method.

6.2. Performance evaluation

For the performance evaluation experiments on text extraction, the parameters for the proposed approach are set according to the optimal results determined in the previous sections. The proposed approach is implemented on a 2.4 GHz Pentium-IV personal computer using C++ programming language. The computation time spent on processing an input document image depends on the size and complexity of the image. Most of the computation time is spent on the multi-plane region matching and assembling process. For a

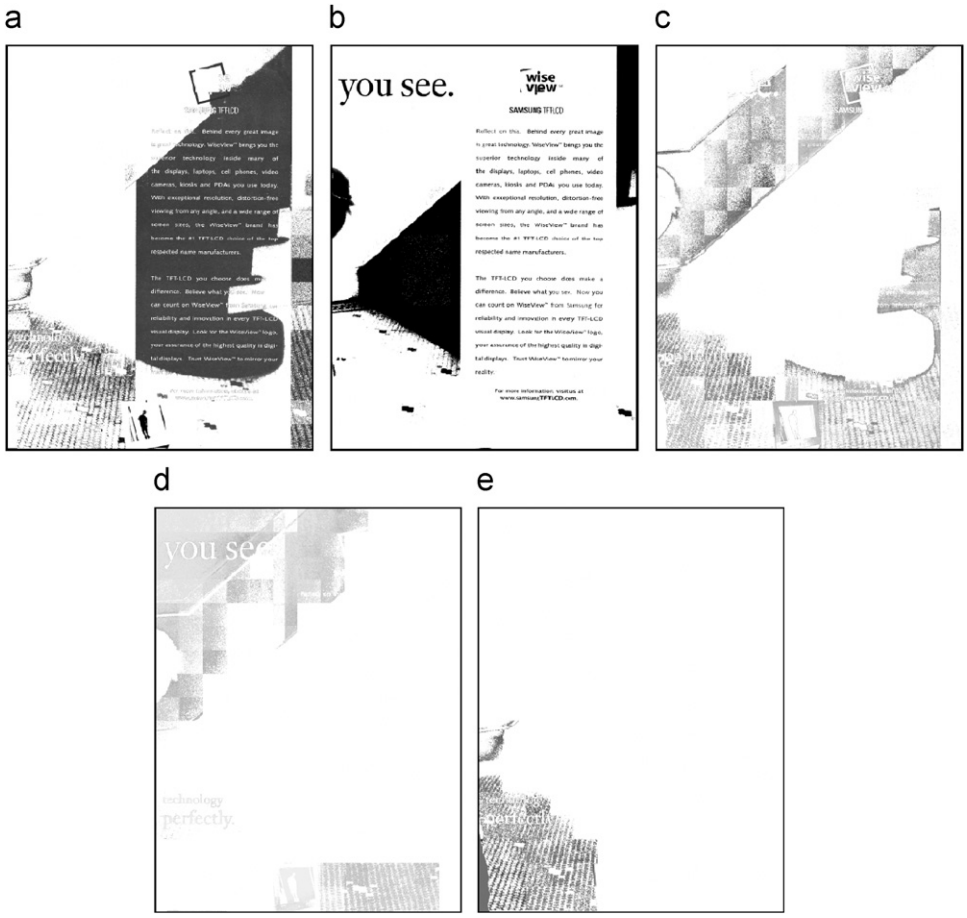

Fig. 14. Decomposed object planes of Fig. 10(b) after performing the proposed multi-plane segmentation: (a) decomposed object plane 1, (b) decomposed object plane 2, (c) decomposed object plane 3, (d) decomposed object plane 4, and (e) decomposed object plane 5.

typical A4-sized document page scanned at 300 dpi resolution, the average image size is 2408 pixels by 3260 pixels, with an average of 1.24 s processing time.

The comparative experiments are firstly conducted on the aforementioned sample image of Fig. 2. Figs. 8 and 9 show the processing results and text extraction results produced by Jain and Yu's color-quantization-based method [24], and Pietikainen and Okun's edge-based method [22]. Here the text extraction results of Pietikainen and Okun's method depicted in Fig. 9(b) and the later figures are converted into masked images where the black mask was adopted to display the non-text regions. As a comparative experiment of document image decomposition, the decomposition results depicted in Figs. 8(a)–(d) are four representative color images after performing Jain and Yu's color quantization method [24]. As can be seen from the second representative color image in Fig. 8(b), the caption characters superimposed on the shaded background are blurred and cannot be appropriately separated. Furthermore, as shown in Fig. 8(d), the bottom text-line "*what now*?" is occluded in the fourth representative color image by reason of the insufficient contrast for color quantization process. As a result, these two textual regions are missed in the resultant text extraction results, as shown in Fig. 9(a). As seen from Fig. 9(b), Pietikainen and Okun's method extracts most characters of the body text, but caption characters are fragmented and characters string "*what now*?" are also lost due to the low contrast with the background.

Figs. 10(a) and (b) are two typical test images of A4 full-size complex scanning documents. The test image shown in Fig. 10(a) contains background objects with sharp illumination variations across textual regions, and some of these also possess similar colors and illuminations to those characters touched with them, so that their illuminations are influenced and have gradational variations due to the scanning process; while the test image in Fig. 10(b) has a large portion of the main body text printed on a large shaded and textured background region, and thus the contrast between the characters and this textured background region is extremely degraded.

Figs. 11 and 12 depict the decomposition results of Fig. 10(a) produced by the proposed multi-plane approach and Jain and Yu's color-quantization-based method, respectively. As shown in Figs. 11(a)–(h), the proposed approach clearly segments the homogenous objects into respective object planes. These planes comprise the textual objects of interest including the large bright characters near the gray boundary blocks in Figs. 11(b) and (e), the characters "SIEMENS" below the man in black in Fig. 11(c), the white main body text close to the mobile phone's shell in Fig. 11(d), and the rest of small characters in Figs. 11(g) and (h). Comparatively, textual objects of the caption and the main body text in Figs. 12(b), (e), and (f) of the representative color images decomposed by Jain and Yu's method are visibly fragmented or blurred with pictorial objects due to the influence of those background objects during the color quantization process. Figs. 13(a)–(c) illustrate the text extraction results of Fig. 10(a) by the proposed approach, Jain and Yu's method, and Pietikainen and Okun's method. As shown in Fig. 13(a), the text extraction results from the proposed approach demonstrate that the

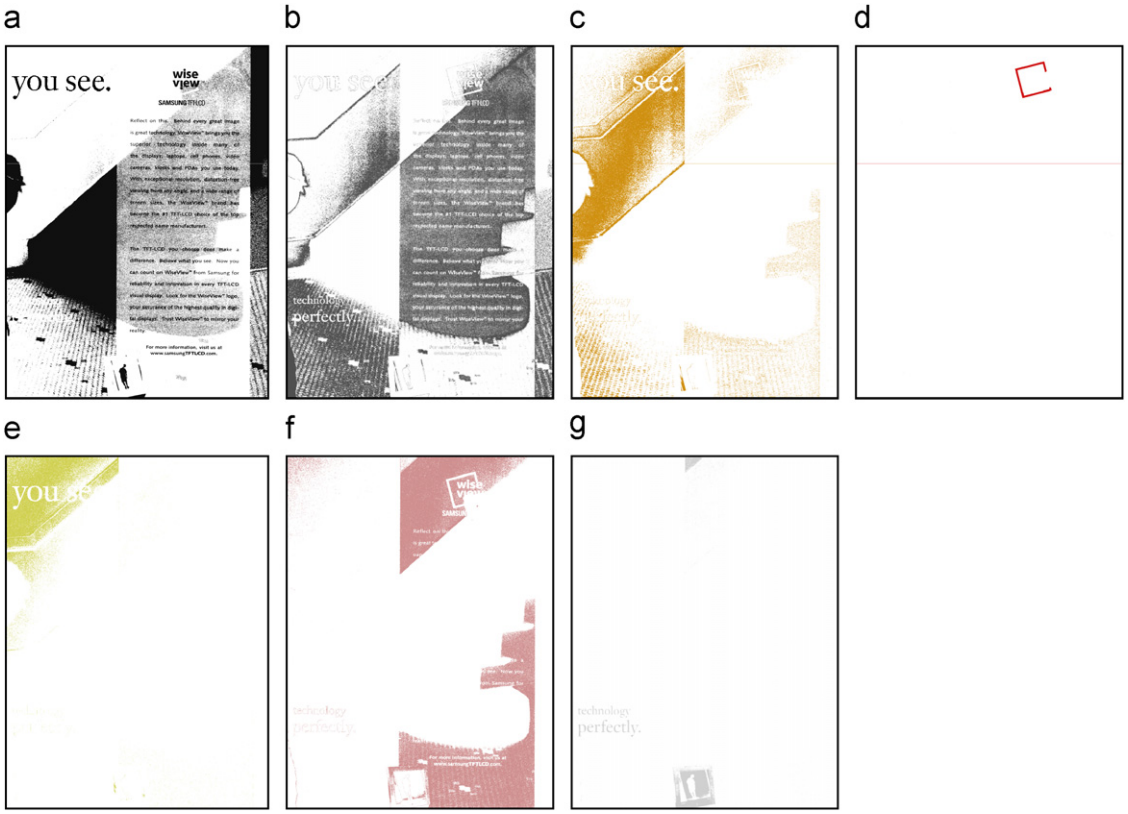

Fig. 15. Representative color images of Fig. 10(b) after performing Jain and Yu's method: (a) representative color image 1, (b) representative color image 2, (c) representative color image 3, (d) representative color image 4, (e) representative color image 5, (f) representative color image 6, and (g) representative color image 7.

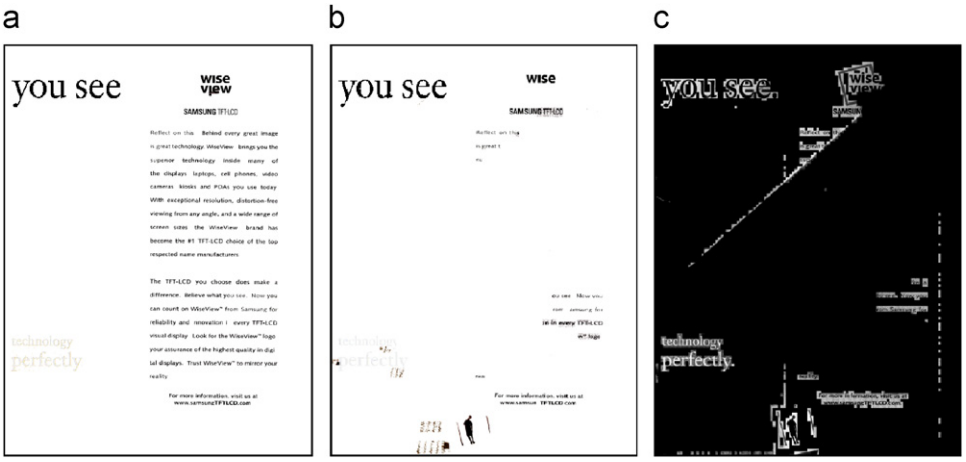

Fig. 16. Text extraction results of Fig. 10(b) by the proposed approach, Jain and Yu's method, and Pietikainen and Okun's method: (a) text extraction results by the proposed approach, (b) text extraction results by Jain and Yu's method, and (c) text extraction results by Pietikainen and Okun's method.

majority of the textual objects are successfully extracted from the sharply varying backgrounds. By comparison, as shown in Fig. 13(b), Jain and Yu's method is unsuccessful in extracting the large caption characters and many characters from the main body of text by reason of the above-mentioned unsatisfactory decomposition results of the color quantization process. Pietikainen and Okun's method extracts most textual objects except some broken large characters and several missed small characters, as shown in Fig. 13(c); however, several pictorial objects with sharp contours are also identified as textual objects, and the characters in extracted textual regions are blurred.

In the test image in Fig. 10(b), the textual regions of interest are the caption characters "you see" on the top-left, the main body of text on the right, and the white characters "technology perfectly" on the bottom-left. Figs. 14–16 illustrate the decomposition and text extraction results on the test image in Fig. 10(b) obtained by the proposed approach, Jain and Yu's method, and Pietikainen and Okun's method, respectively. As shown in Fig. 14(b), the proposed approach correctly separates the main body of text printed on shaded and textured background regions in highly degraded contrasts. By comparison, textual regions of the main body of text in Fig. 15(a) of the representative color image obtained by Jain and Yu's method are

Fig. 17. Original images of the test images 4–6: (a) test image 4 (size: 2864×3658), (b) test image 5 (size: 2427×3166), and (c) test image 6 (size: 2469×3535).

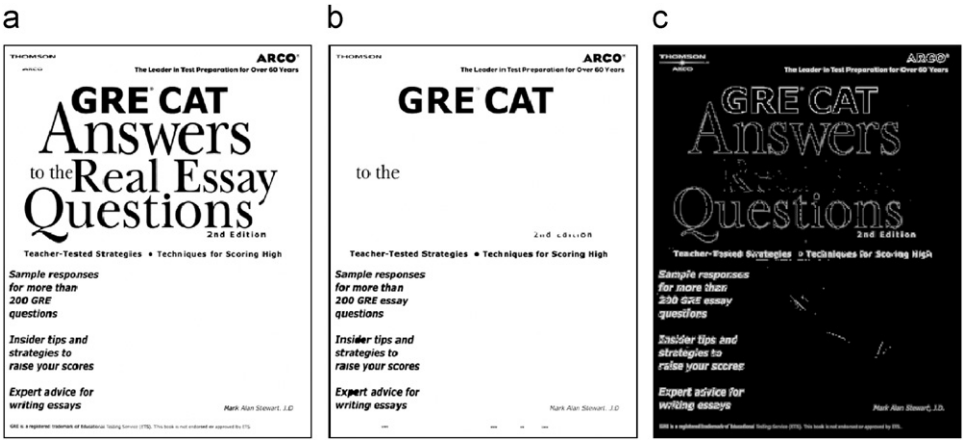

Fig. 18. Text extraction results of Fig. 17(a) by the proposed approach, Jain and Yu's method, and Pietikainen and Okun's method: (a) binarized text extraction results by the proposed approach, (b) binarized text extraction results by Jain and Yu's method, and (c) text extraction results by Pietikainen and Okun's method.

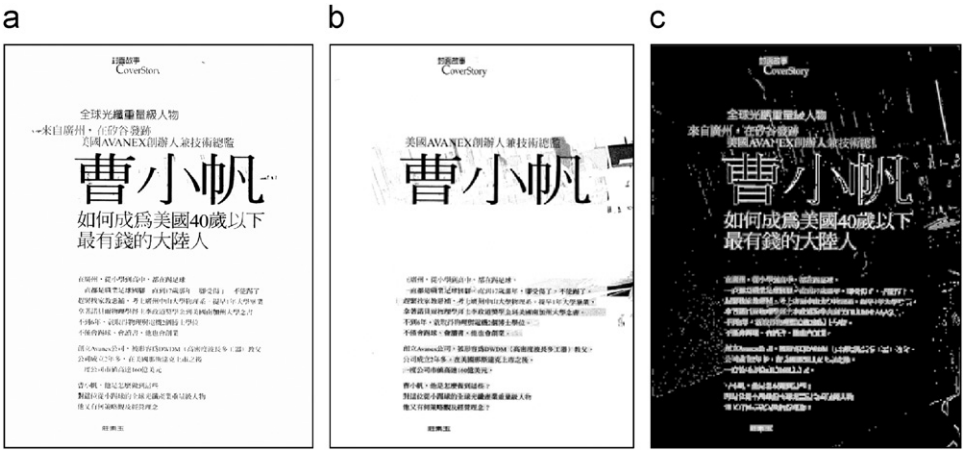

Fig. 19. Text extraction results of Fig. 17(b) by the proposed approach, Jain and Yu's method, and Pietikainen and Okun's method: (a) binarized text extraction results by the proposed approach, (b) binarized text extraction results by Jain and Yu's method, and (c) text extraction results by Pietikainen and Okun's method.

smeared with the background regions. Accordingly, as can be seen from Fig. 16(a), the characters in three different textual regions are successfully extracted by the proposed approach; whereas both work of Jain and Yu, and Pietikainen and Okun perform not so well on extracting textual objects from the shaded and textured backgrounds in degraded contrasts, as shown in Figs. 16(b) and (c), respectively.

Figs. 17(a)–(c) are three test images with several notable characteristics. The test image in Fig. 17(a) has multiple-colored

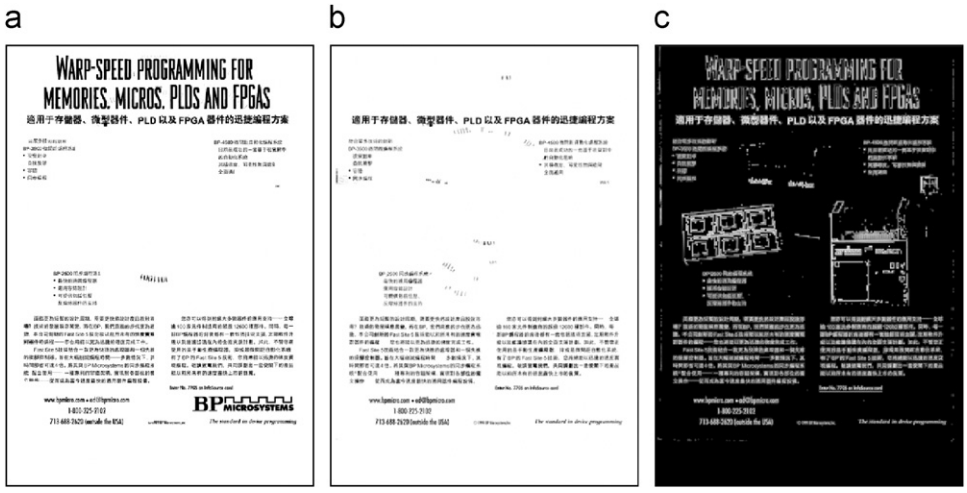

Fig. 20. Text extraction results of Fig. 17(c) by the proposed approach, Jain and Yu's method, and Pietikainen and Okun's method: (a) binarized text extraction results by the proposed approach, (b) binarized text extraction results by Jain and Yu's method, and (c) text extraction results by Pietikainen and Okun's method.

Table 5
Experimental data of Jain and Yu's method and our approach on the English test database.

Method	Recall rate (%)	Precision rate (%)
Jain and Yu's method	85.1	98.7
Our approach	99.4	99.6

Table 6
Experimental data of Jain and Yu's method and our approach on the Chinese test database.

Method	Recall rate (%)	Precision rate (%)
Jain and Yu's method	82.1	95.2
Our approach	99.2	99.4

text-lines printed on several shaded background regions in indistinct contrasts, while the test images in Figs. 17(b) and (c) comprise textual regions overlapped with numerous character-like objects with similar contrasts and textural features to those of actual textual objects. To facilitate the visual observation of bright characters, the text extraction results of Jain and Yu's method and the proposed approach in Figs. 18(a) and (b), 19(a) and (b), and 20(a) and (b) are illustrated in the binarized form. Figs. 18(a), 19(a), and 20(a) exhibit that the proposed approach correctly segments and extracts the textual objects with different sizes, types, and colors under various difficulties associated with the complexity of background images. As shown in Figs. 18(b), 19(b), and 20(b) Jain and Yu's method could not perform well on extracting several text-lines of interest, and some extracted textual regions are also blurred or degraded. As illustrated in Figs. 18(c), 19(c), and 20(c), Pietikainen and Okun's method can extract most textual objects, but several shaded textual objects such as the caption characters "to the Real Essay" in Fig. 18(c) are still missed, and many background textures and contoured objects are also identified as textual objects, and so that many extracted textual regions are blotted by these spurious detections.

To perform the quantitative evaluation of text extraction performance, two evaluation criteria, including the above-mentioned recall rate (as described in Eq. (37)), and the precision rate as defined in Eq. (38), which are commonly used for evaluating performance in information retrieval, are adopted.

$$\text{Precision rate} = \frac{\text{No. of correctly extracted characters}}{\text{No. of extracted character-like components}} \quad (38)$$

Similar to the above-mentioned definition of the recall rate, the precision rate for text extraction results is obtained by manually counting the number of total extracted character-like connected-components and the correctly extracted characters from the document image, respectively. The experiments of quantitative evaluation were performed on our English and Chinese test databases, the English test database contains 28 English document images with totaling 20,391 visible characters, while the Chinese test database contains 37 Chinese and mixed-Chinese/English document images with totaling 19,635 visible characters. From the text extraction viewpoint, the recall rate reveals the percentage of correctly extracted characters as opposed to all actual characters within each processed document image, while the precision rate represents the percentage of correctly extracted characters as opposed to all extracted character-like connected-components. Since these quantitative evaluation criteria are performed on the extracted connected-components, the results of Pietikainen and Okun's method are inappropriate for evaluation using these criteria, and were not involved in the quantitative evaluation.

Tables 5 and 6, respectively, depict the results of quantitative evaluation on the English and Chinese test databases of Jain and Yu's method and the proposed approach. As for the computation timing issue, the average computing times of the Jain and Yu's method and the proposed approach on processing an A4-sized document image with 300 dpi scanned resolution are 0.88 and 1.24 s using the above-mentioned platform, respectively. Therefore, by Tables 5 and 6, and the computational timings, it is observed that although the proposed approach costs a little computing time, the proposed approach can provide better text extraction performance as compared to that of Jain and Yu's method.

Figs. 21–26 show several further examples of the proposed approach on extracting English and Chinese textual objects from complex document images. Although a few non-text components with character-like characteristics are detected as textual objects, and a few small punctuation marks are missed because of their small sizes and non-alignment with other characters contained in text-lines, the overwhelming majority of the textual objects are correctly obtained.

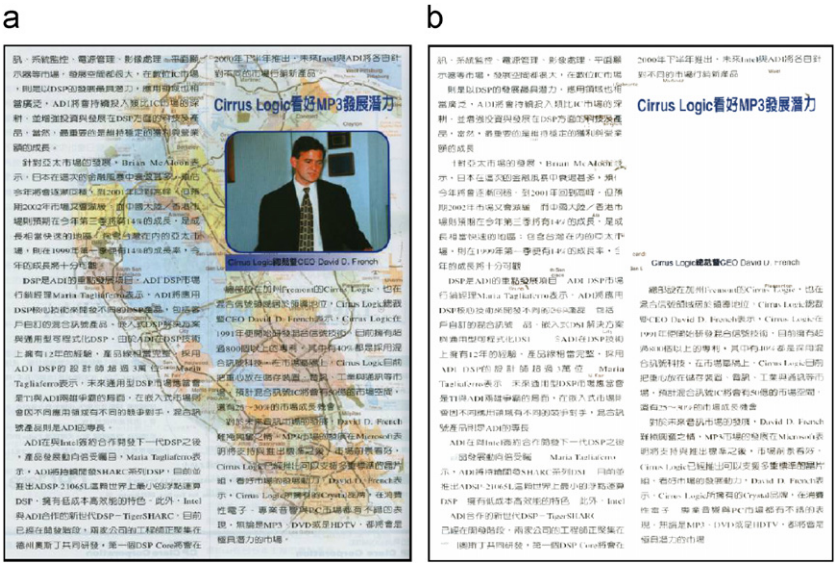

Fig. 21. Results of test image 7 (size: 1829×2330): (a) original image and (b) text extraction results by the proposed approach.

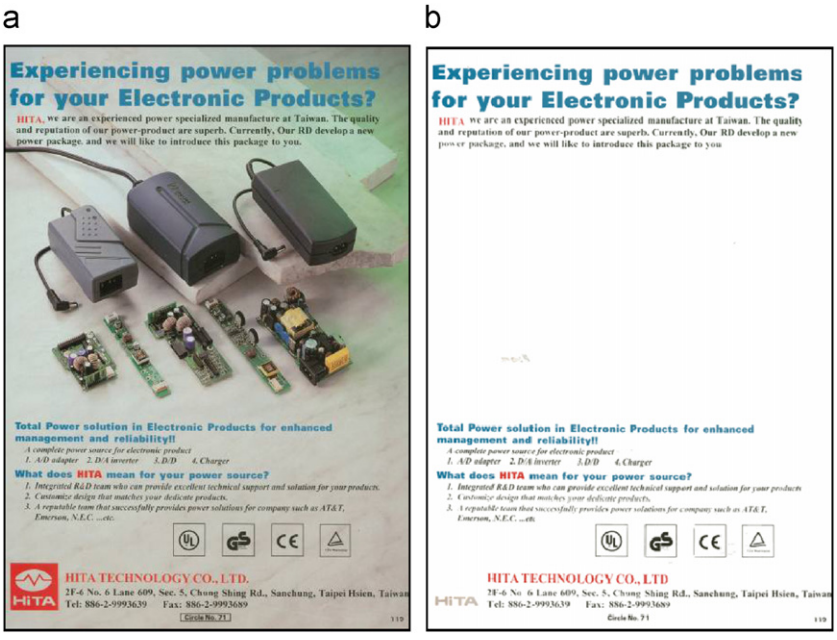

Fig. 22. Results of test image 8 (size: 3147×4536): (a) original image and (b) text extraction results by the proposed approach.

By observing the results obtained, even if textual objects comprised of various illuminations, sizes, font types, and styles, are overlapped with pictorial objects and backgrounds with uneven, gradational, and sharp variations in contrast, illumination, and texture, almost all the textual objects are effectively detected and extracted by the proposed approach.

7. Conclusions

A new technique for segmenting and extracting textual objects from real-life complex document images is presented in this study. The proposed approach first segregates textual regions, non-text objects such as graphics and pictures, and background textures from the document image by decomposing it into distinct object planes. This decomposition process consists of two stages: automatic localized histogram multilevel thresholding and multi-plane region matching and assembling. The first stage applies the localized histogram multilevel thresholding procedure to discriminate different textual objects, non-textual objects, and background components in each block region into separate *SR*s. In the second stage, the multi-plane region matching and assembling process organizes these obtained *SR*s into object planes according to their respective features. A text extraction procedure is then applied to the resultant planes to extract textual objects with different characteristics in the corresponding planes. The document image is processed regionally and adaptively according to its local features, and thus detailed

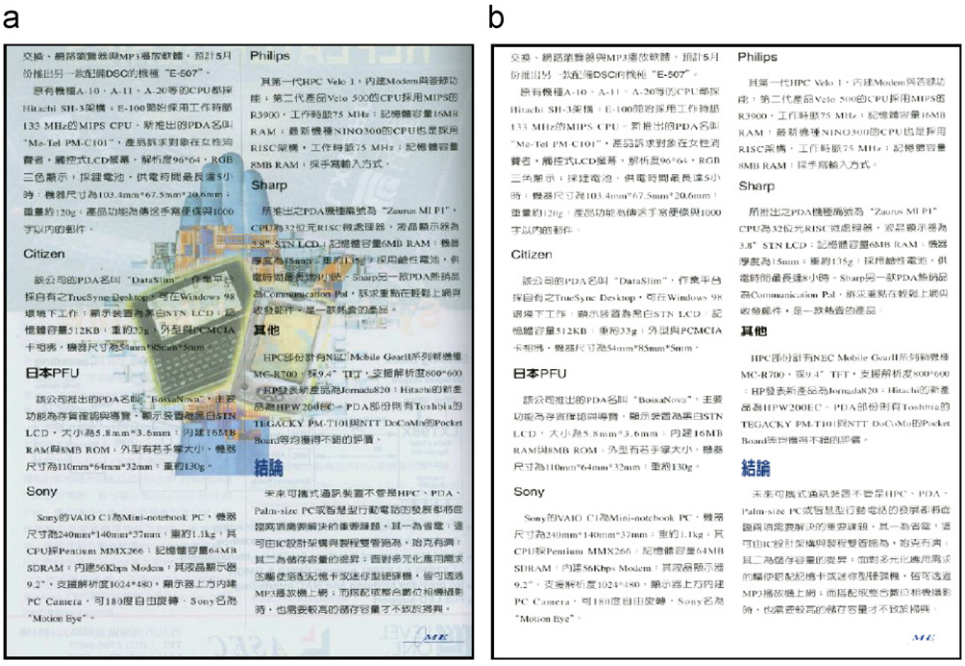

Fig. 23. Results of test image 9 (size: 1859×2437): (a) original image and (b) text extraction results by the proposed approach.

Fig. 24. Results of test image 10 (size: 1344×1792): (a) original image and (b) text extraction results by the proposed approach.

characteristics of extracted textual objects can be well-preserved, especially small characters with thin strokes. It also allows textual objects that touch graphical and pictorial background objects with uneven, gradational, and sharp variations in contrast, illumination, and texture to be well-handled. When applied to real-life complex document images, the proposed approach exhibits its robustness in extracting horizontal textual objects of various illuminations, sizes, and font styles from complex backgrounds. From the experimental results and comparisons to other existing works, the proposed approach demonstrates its effectiveness and advantages for most real-life complex document images. However, the text extraction method of the current study could just accommodate some small skews of scanned document images and text lines. Therefore, in the further studies, the text extraction method can be improved by integrating an effective skew estimation and correction technique, such as the effective techniques presented in Refs. [27,43–45], to efficiently handle skewed document images and text lines with different orientations.

Acknowledgments

The authors would like to thank the reviewers for their valuable comments, which have helped to improve the manuscript. This work was supported by the National Science Council of R.O.C. under Contract No. 96-2218-E-468-006 and Asia University under Contract No. 97-I-01.

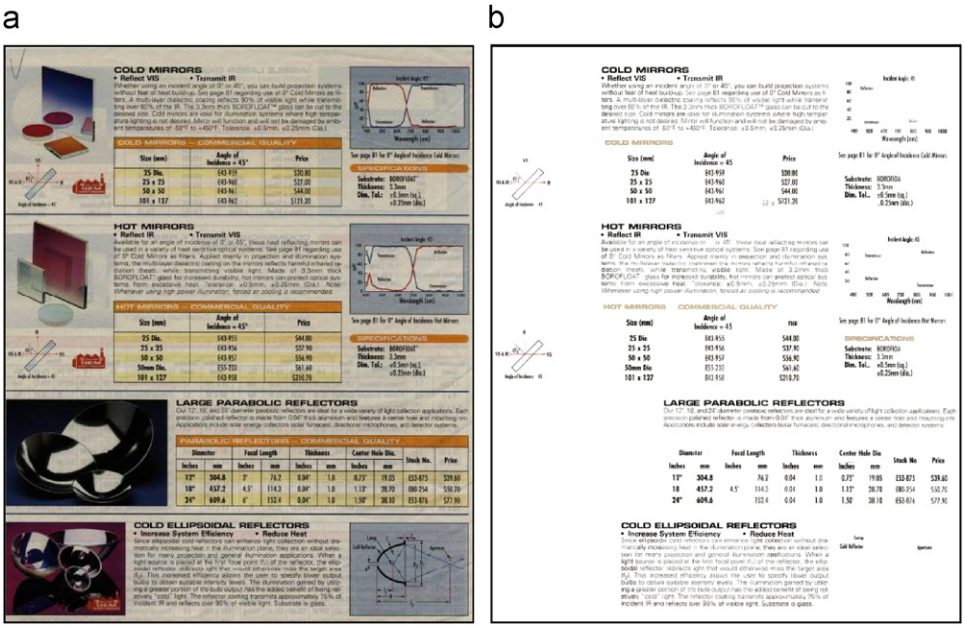

Fig. 25. Results of test image 11 (size: 2309×2829): (a) original image and (b) text extraction results by the proposed approach.

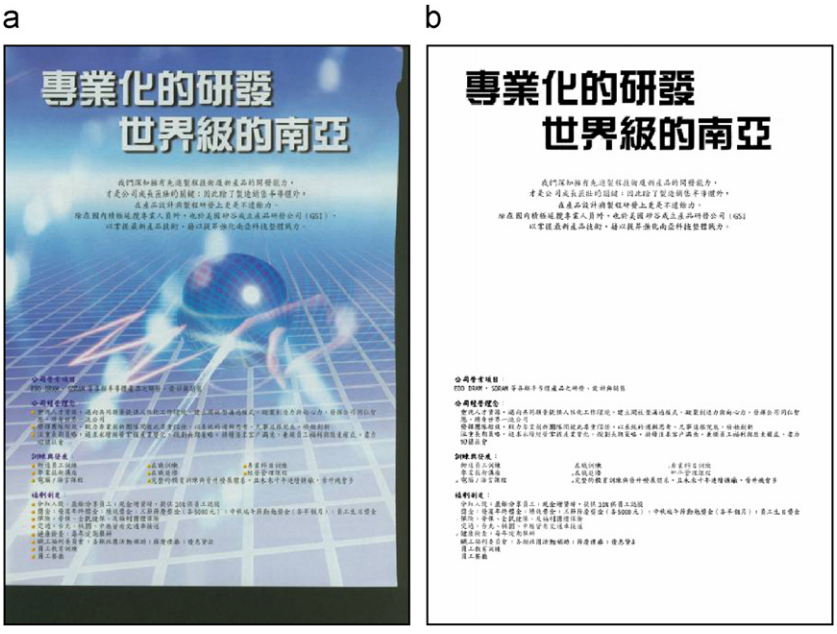

Fig. 26. Results of test image 12 (size: 2469×3535): (a) original image and (b) binarized text extraction results by the proposed method.

References

[1] D. Doermann, The indexing and retrieval of document images: a survey, Comput. Vision Image Understanding 70 (1998) 287–298.

[2] in: H. Bunke, P.S.P. Wang (Eds.), Handbook of Character Recognition and Document Image Analysis, World Scientific, Singapore, 1997.

[3] L. O' Gorman, R. Kasturi, Document Image Analysis, IEEE Computer Society Press, Silver Spring, MD, 1995.

[4] L.A. Fletcher, R. Kasturi, A robust algorithm for text string separation from mixed text/graphics images, IEEE Trans. Pattern Anal. Mach. Intell. 10 (6) (1988) 910–918.

[5] J.L. Fisher, S.C. Hinds, D.P. D'Amato, Rule-based system for document image segmentation, in: Proceedings of the 10th International Conference on Pattern Recognition, 1990, pp. 567–572.

[6] F.Y. Shih, S.S. Chen, D.C.D. Hung. P.A. Ng, Document segmentation, classification and recognition system, in: Proceedings of the IEEE International Conference on Systems Integration, 1992, pp. 258–267.

[7] Y. Liu, S.N. Srihari, Document image binarization based on texture features, IEEE Trans. Pattern Anal. Mach. Intell. 19 (5) (1997) 540–544.

[8] M. Cheriet, J.N. Said, C.Y. Suen, A recursive thresholding technique for image segmentation, IEEE Trans. Image Process. 7 (6) (1998) 918–921.

[9] N. Otsu, A threshold selection method from gray-level histograms, IEEE Trans. Syst. Man Cybern. SMC-8 (1978) 62–66.

[10] Y. Solihin, C.G. Leedham, Integral ratio: a new class of global thresholding techniques for handwriting images, IEEE Trans. Pattern Anal. Mach. Intell. 21 (8) (1999) 761–768.

[11] J.R. Parker, Gray level thresholding in badly illuminated images, IEEE Trans. Pattern Anal. Mach. Intell. 13 (8) (1991) 813–819.

[12] J. Ohya, A. Shio, S. Akamatsu, Recognizing characters in scene images, IEEE Trans. Pattern Anal. Mach. Intell. 16 (2) (1994) 214–220.

[13] M. Kamel, A. Zhao, Extraction of binary character/graphics images from grayscale document images, CVGIP: Graphical Models Image Process. 55 (3) (1993) 203–217.

[14] N.B. Venkateswarlu, R.D. Boyle, New segmentation techniques for document image analysis, Image and Vision Comput. 13 (7) (1995) 573–583.
[15] Y. Yang, H. Yan, An adaptive logical method for binarization of degraded document images, Pattern Recognition 33 (2000) 787–807.
[16] X. Ye, M. Cheriet, C.Y. Suen, Stroke-model-based character extraction from gray-level document images, IEEE Trans. Image Process. 10 (8) (2001) 1152–1161.
[17] A. Dawoud, M. Kamel, Iterative multi-model sub-image binarization for handwritten character segmentation, IEEE Trans. Image Process. 13 (9) (2004) 1223–1230.
[18] A. Amin, S. Wu, A robust system for thresholding and skew detection in mixed text/graphics documents, Int. J. Image Graphics 5 (2) (2005) 247–265.
[19] Q. Yuan, C.L. Tan, Text extraction from gray scale document images using edge information, in: Proceedings of the Sixth International Conference on Document Analysis and Recognition, 2001, pp. 302–306.
[20] V. Wu, R. Manmatha, E.M. Riseman, Textfinder: an automatic system to detect and recognize text in images, IEEE Trans. Pattern Anal. Mach. Intell. 21 (11) (1999) 1224–1229.
[21] Y.M.Y. Hasan, L.J. Karam, Morphological text extraction from images, IEEE Trans. Image Process. 9 (11) (2000) 1978–1983.
[22] M. Pietikinen, O. Okun, Edge-based method for text detection from complex document images, in: Proceedings of the Sixth International Conference on Document Analysis and Recognition, 2001, pp. 286–291.
[23] Y. Zhong, K. Karu, A.K. Jain, Locating text in complex color images, Pattern Recognition 28 (10) (1995) 1523–1535.
[24] A.K. Jain, B. Yu, Automatic text location in images and video frames, Pattern Recognition 31 (12) (1998) 2055–2076.
[25] C. Strouthopoulos, N. Papamarkos, A.E. Atsalakis, Text extraction in complex color documents, Pattern Recognition 35 (2002) 1743–1758.
[26] H. Yang, S. Ozawa, Extraction of bibliography information based on the image of book cover, IEICE Trans. Inf. Syst. E 82-D (7) (1999) 1109–1116.
[27] H. Hase, M. Yoneda, S. Tokai, J. Kato, C.Y. Suen, Color segmentation for text extraction, Int. J. Doc. Anal. Recognition 6 (4) (2004) 271–284.
[28] B.-F. Wu, C.-C. Chiu, Y.-L. Chen, Compound document compression algorithms for text/background overlapping images, IEE Proc. Vision Image Signal Process. 151 (6) (2004) 453–459.
[29] O.D. Trier, T. Taxt, Evaluation of binarization methods for document images, IEEE Trans. Pattern Anal. Mach. Intell. 17 (1995) 312–314.
[30] O.D. Trier, A.K. Jain, Goal-directed evaluation of binarization methods, IEEE Trans. Pattern Anal. Mach. Intell. 17 (1995) 1191–1201.
[31] B.-F. Wu, Y.-L. Chen, C.-C. Chiu, A discriminant analysis based recursive automatic thresholding approach for image segmentation, IEICE Trans. Inf. Syst. E 88-D (7) (2005) 1716–1723.
[32] R. Kasturi, M.M. Trivedi, Image Analysis Applications, Marcel Dekker, New York, 1990.
[33] A. Rosenfeld, A.C. Kak, Digital Picture Processing, second ed., vol. 2, Academic Press, New York, 1982.
[34] R.M. Haralick, L.G. Shapiro, Computer and Robot Vision, vol. I, Addison-Wesley, Reading, MA, 1992.
[35] R.R. Yager, D.P. Filev, Approximate clustering via the mountain method, IEEE Trans. Syst. Man Cybern. 24 (8) (1994) 1279–1284.
[36] S.L. Chiu, Extracting fuzzy rules for pattern classification by cluster estimation, in: Proceedings of the Sixth International Fuzzy Systems Association World Congress, 1995, pp. 1–4.
[37] N.R. Pal, D. Chakraborty, Mountain and subtractive clustering method: improvements and generalization, Int. J. Intell. Syst. 15 (2000) 329–341.
[38] K. Suzuki, I. Horiba, N. Sugie, Linear-time connected-component labeling based on sequential local operations, Comput. Vision Image Understanding 89 (2003) 1–23.
[39] J. Ha, R.M. Haralick, I. Phillips, Document page decomposition by the bounding-box projection technique, in: Proceedings of the Third International Conference on Document Analysis and Recognition, 1995, pp. 1119–1122.
[40] J. Ha, R.M. Haralick, I. Phillips, Recursive *X*–*Y* cut using bounding boxes of connected components, in: Proceedings of the Third International Conference on Document Analysis and Recognition, 1995, pp. 952–955.
[41] T. Pavlidis, J. Zhou, Page segmentation and classification, Comput. Vision Graphical Image Process. 54 (6) (1992) 484–496.
[42] F.Y. Shih, S.S. Chen, Adaptive document block segmentation and classification, IEEE Trans. Syst. Man Cybern. B 26 (5) (1996) 797–802.
[43] B. Yu, A. Jain, A robust and fast skew detection algorithm for generic documents, Pattern Recognition 29 (10) (1996) 1599–1629.
[44] B. Gatos, N. Papamarkos, C. Chamzas, A Fast algorithm for skew detection and text line position determination in digitized documents, Pattern Recognition 30 (9) (1997) 1505–1519.
[45] C.H. Chou, S.Y. Chu, F. Chang, Estimation of skew angles for scanned documents based on piecewise covering by parallelograms, Pattern Recognition 40 (2) (2007) 443–455.

About the Author—YEN-LIN CHEN was born in Kaohsiung, Taiwan in 1978. He received the B.S. and Ph.D. degree in electrical and control engineering from National Chiao Tung University, Hsinchu, Taiwan, in 2000 and 2006, respectively. Since 2007, he is an Assistant Professor at the Department of Computer Science and Information Engineering, Asia University, Taichung, Taiwan.
Dr. Chen is a member of the IEEE, IAPR, and IEICE. In 2003, he received Dragon Golden Paper Award sponsored by the Acer Foundation and the Silver Award of Technology Innovation Competition sponsored by the AdvanTech. His research interests include image and video processing, pattern recognition, document image analysis, and intelligent transportation system.

About the Author—BING-FEI WU was born in Taipei, Taiwan in 1959. He received the B.S. and M.S. degrees in control engineering from National Chiao Tung University (NCTU), Hsinchu, Taiwan, in 1981 and 1983, respectively, and the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, in 1992.
From 1983 to 1984, he was with the Institute of Control Engineering, NCTU as an Assistant Researcher. From 1985 to 1988, he was with the Department of Communication Engineering at the same university as a Lecturer. Since 1992, he has been with the Department of Electrical Engineering and Control Engineering, where he is currently a Professor. As an active industry consultant, he is also involved in the chip design and applications of the flash memory controller and 3C consumer electronics in multimedia. His research interests include chaotic systems, fractal signal analysis, multimedia coding, and wavelet analysis and applications.