2006 IEEE International Conference on
Systems, Man, and Cybernetics
October 8-11, 2006, Taipei, Taiwan

Text Extraction from Complex Document Images Using the Multi-plane Segmentation Technique

Yen-Lin Chen, *Student Member, IEEE*, and Bing-Fei Wu*, *Senior Member, IEEE*

Abstract—This study presents a new method for extracting characters from various real-life complex document images. The proposed method applies a multi-plane segmentation technique to separate homogeneous objects including text blocks, non-text graphical objects, and background textures into individual object planes. It consists of two stages - automatic localized multilevel thresholding, and multi-plane region matching and assembling. Then a text extraction process can be performed on the resultant planes to detect and extract characters with different characteristics in the respective planes. The proposed method processes document images regionally and adaptively according to their respective local features. This allows preservation of detailed characteristics from extracted characters, especially small characters with thin strokes, as well as gradational illuminations of characters. This also permits 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 method is effective in extracting characters with various illuminations, sizes, and font styles from various types of complex document images.[1]

I. Introduction

Textual information extraction from document images provides many useful applications in document analysis and understanding, such as optical character recognition, document retrieval, and compression. To-date, many effective techniques have been developed for extracting characters from monochromatic document images [1]. In recent years, advances in multimedia publishing and printing technology have led to an increasing number of real-life documents in which stylistic text strings are printed with decorated objects and colorful, varied background components. However, these approaches do not work well for extracting characters from real-life complex document images. Compared to monochromatic document images, text extraction in complex document images brings with it many difficulties associated with the complexity of background images, variety and shading of character illuminations, superimposing characters with illustrations and pictures, as well as other decorated background components.

Since most characters show sharp and distinctive edge features, methods based on edge information [2]-[4] have been developed. Such methods utilize an edge detection operator to extract the edge features of character objects, and then use these features to extract characters from document images. Such edge-based methods are capable of extracting characters in different homogeneous illuminations from graphic backgrounds. However, when the characters 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 text, and thus the characters in extracted text regions are blurred by those non-text objects.

In recent years, several color segmentation-based methods for text extraction from color document images have been proposed [5]-[7]. These methods utilize color clustering or quantization approaches for determining the prototype colors of documents so as to facilitate the detection of character objects in these separated color planes. However, most of these methods have difficulties in extracting characters which are embedded in complex backgrounds or that touch other graphical objects. This is because the prototype colors are determined in a global view, so that appropriate prototype colors cannot be easily selected for distinguishing characters from those touched graphical objects and complex backgrounds without sufficient contrast.

In this study, we propose an effective method for extracting characters from these complex document images, and resolving the above issues associated with the complexity of their backgrounds. The document image is processed by the proposed multi-plane segmentation technique to decompose it into separate object planes. The proposed multi-plane segmentation technique comprises two stages: automatic localized histogram multilevel thresholding, and multi-plane region matching and assembling processing. After the multi-plane segmentation technique has been carried out, homogeneous objects including text blocks, other non-text objects, and background textures are separated into individual object planes. The text extraction process is then performed on the resultant planes to detect and extract characters with different characteristics in the respective planes. The document image is processed regionally and adaptively according to local features by the proposed method. This allows detailed characteristics of the extracted characters to be well-preserved, especially the small characters with thin strokes, as well as the gradational illuminations of characters. This also allows for characters adjoined or touched with graphical objects and backgrounds with uneven, gradational, and sharp variations in contrast, illumination, and texture to be handled easily and well. Experimental results demonstrate that the proposed method is capable of extracting characters from various types of complex document images.

II. Localized Histogram Multilevel Thresholding

The multi-plane segmentation, if necessary, begins by applying a color-to-grayscale transformation on the *RGB* components of a color document image to obtain its illumination component *Y*. After the color transformation is performed, the illumination image *Y* still retains the texture features of the original color image, as pointed out in [2], and

[1]This work is supported by the National Science Council under Grant No. NSC 94-2213-E-009-066.
*The authors are with Department of Electrical and Control Engineering, National Chiao Tung University, 1001 Ta-Hsueh Road, Hsinchu 30050, Taiwan (e-mail: bwu@cc.nctu.edu.tw)

thus the character strokes in their original color are still well-preserved. Then the obtained illumination image Y will be divided into non-overlapping rectangular block regions with dimension $M_H \times M_V$, as shown in Fig. 1(b). Thus the mission is to extract objects with similar characteristics from these rectangular block regions into different sub-block regions to facilitate further analysis in the following stage. Hence, an effective multilevel thresholding technique is needed for automatically determining the suitable number of thresholds for segmenting 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 [8]. This technique automatically determines the suitable number of thresholds, and utilizes a fast recursive selection strategy for selecting the optimal thresholds to segment the image into separate objects with similar characteristics in a computationally frugal way. In this study, we utilize this multilevel threshold selection technique and make necessary modifications to adapt it for segmenting block regions into different objects with similar characteristics.

We utilize the concept of separability measure described in [8] as a criterion of automatic segmentation of objects in a given block region, denoted by $\Re$. This segmentation criterion is denoted by the "separability factor" – $\mathcal{SF}$ as in [8], and is defined as,

$$\mathcal{SF} = \nu_{BC}(\boldsymbol{T})/\nu_{\Re} = 1-\nu_{WC}(\boldsymbol{T})/\nu_{\Re} \qquad (1)$$

where ν_{BC}, ν_{WC} and $\nu_{\Re}$ are between-class variance, within-class variance, and total variance of the gray intensities of pixels in the region $\Re$, and $\nu_{\Re}$ serves as the normalization factor; and $\boldsymbol{T}$ is the evaluated threshold set composed of n thresholds, to partition pixels in the region $\Re$ into $n+1$ classes. These pixel classes are represented by $C_0 = \{0,...,t_1\},\ldots, C_k = \{t_k+1,...,t_{k+1}\},\ldots, C_n = \{t_n+1,...,U-1\}$.

Based on this efficient discriminant criterion, an automatic multilevel thresholding is applied for recursively segmenting the block region $\Re$ into different objects of homogeneous illuminations, regardless of the number of objects and image complexity of the region $\Re$. It can be performed until the $\mathcal{SF}$ measure is large enough to show that the appropriate discrepancy among the resultant classes has been obtained. This objective is reached by the scheme that selects the class with the maximal contribution $w_k\sigma_k^2$ (where w_k and σ_k^2 are the cumulative probability and variance of the gray intensities in class C_k, respectively) of the total within-class variance $\nu_{WC}(\boldsymbol{T})$, denoted by C_p, for recursively performing an optimal *bi-class partition procedure*, as described in [8], until the separability among all classes becomes satisfactory, i.e. the condition where the $\mathcal{SF}$ 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 thresholding procedure as described in [8]. Thus, the $\mathcal{SF}$ 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 illuminations will be well-separated.

Furthermore, if a region $\Re$ comprises of a set of pixels with homogeneous gray intensity features, consisting of parts of a larger object or background region, then it should not be partitioned and should keep its original components. These homogeneous regions can be determined by evaluating two statistical features: 1) the bi-class $\mathcal{SF}$ measure, denoted as $\mathcal{SF}_b$, which is the $\mathcal{SF}$ value obtained by performing the initial *bi-class partition procedure* on region $\Re$, i.e. the $\mathcal{SF}$ value associated with the determined threshold t_S^*; and 2) the illumination variance, $\nu_{\Re}$ of pixels in the region $\Re$. Hence, if both the $\mathcal{SF}_b$ and $\nu_{\Re}$ features have small values, this reveals that the distribution of the region $\Re$ is concentrated within a compact range, and thus the $\Re$ comprises a set of homogeneous pixels representing a simple object or parts thereof. Therefore, the following *homogeneity condition* is utilized for determining the situation where both the $\mathcal{SF}_b$ and $\nu_{\Re}$ features are small:

$$\mathcal{SF}_b < Th_{h0} \text{ , and } \nu_{\Re} < Th_{h1} \qquad (2)$$

where Th_{h0} and Th_{h1} are pre-defined thresholds. Therefore, if the homogeneity condition is satisfied, the region $\Re$ is recognized as a homogeneous region, and does not need to undergo the partition process and hence keeps its pixels of homogeneous objects unchanged to be processed by the next stage. The values of the two thresholds Th_{h0} and Th_{h1} are experimentally chosen as 0.6 and 90, respectively.

Then the localized automatic multilevel thresholding process is performed as the following steps:

Step 1: To begin, the illumination image Y with size $W_{img} \times H_{img}$ is divided into rectangular block regions $\Re^{i,j}$ with dimension $M_H \times M_V$, as shown in Fig. 1(b). Here (i, j) are the location indices, and $i = 0,..., N_H$ and $j = 0,..., N_V$, where $N_H = \left(\lceil W_{img}/M_H \rceil - 1\right)$ and $N_V = \left(\lceil H_{img}/M_V \rceil - 1\right)$, 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 the illumination variance - $\nu_{\Re}^{i,j}$ and the bi-class separability measure $\mathcal{SF}_b$; 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. (2), 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 pixels decomposed from $\Re^{i,j}$, where k denotes the index of the present classes and $n = 0, \ldots, q\text{-}1$.

Step 4: From all classes $C_k^{i,j}$, determine the class $C_p^{i,j}$ which has the maximal contribution ($w_k^{i,j}\sigma_k^{i,j\,2}$) 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 $\mathcal{SF}$.

Step 5: Partition $C_p^{i,j}:\{t_p^{i,j}+1, ..., t_{p+1}^{i,j}\}$ into two classes $C_{p0}^{i,j}:\{t_p^{i,j}+1, ..., t_S^{i,j}{}^*\}$, and $C_{p1}^{i,j}:\{t_S^{i,j}{}^*+1, ..., t_{p+1}^{i,j}\}$, 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, $C_0^{i,j}$, ..., $C_{p0}^{i,j}$, $C_{p1}^{i,j}$, ..., $C_{q-1}^{i,j}$, and then let $q = q+1$ to update the record of the current class amount.

Step 6: Compute the $\mathcal{SF}$ value of all currently obtained classes using Eq. (1), if the *objective condition*, $\mathcal{SF} \geq Th_{SF}$, 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 sub-block regions, $SR^{i,j,0}$, $SR^{i,j,1}$, ..., $SR^{i,j,q-1}$, corresponding to the partitioned classes of gray illumination intensities, $C_0^{i,j}$, $C_1^{i,j}$, ..., $C_{q-1}^{i,j}$, respectively, where the notation $SR^{i,j,k}$ represents the k-th *SR* decomposed from the region $\Re^{i,j}$. Then, finish the localized thresholding process on $\Re^{i,j}$ and go back to step 2 and repeat Steps 2~6 to process 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 sub-block regions of the corresponding region $\Re$.

The value of the separability measure threshold Th_{SF} is chosen as 0.92 according to the experimental analysis described in [8] to yield satisfactory segmentation results on the block regions. With regard to the dimension $M_H \times M_V$ of the block region, suitable larger values of the parameters M_H and M_V should be selected so that all foreground objects in the images can be clearly segmented. Since a typical resolution of the document images ranges from 200 dpi to 600 dpi for scanning books, advertisements, journals and magazines, etc, we utilize $M_H = M_V = 100$ (8.5 mm on the 300 dpi resolution) in our experiments. The multi-plane region matching and assembling process, as presented in the following section, is then carried out to assemble and classify them into object planes, denoted as $\mathcal{P}$. All *SR*s obtained from the localized multilevel thresholding process are collected into a hypothetical "*Pool*", in which the multi-plane region matching and assembling process is conducted.

III. Multi-plane Region Matching and Assembling

This section describes an algorithm for constructing the object planes from the *SR*s, generated from the localized multilevel thresholding procedure introduced in the preceding section. Some statistical and spatial features of the adjacent *SR*s are introduced into the multi-plane region matching and assembling procedure in order to assemble all *SR*s of the homogenous text region or object. The proposed multi-plane region matching and assembling process is conducted by performing the following three phases – the *initial plane selection phase*, the *matching phase* and the *plane constructing decision phase*.

Several concepts and definitions are first introduced to facilitate the matching and assembling process of *SR*s obtained from the previous localized thresholding procedure. An *SR* may comprise several connected object regions of pixels decomposed from its associated region $\Re$. Thus the pixels that belong to the 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 one *SR* is defined as follows,

$$\boldsymbol{OP}(SR^{i,j,k}) = \{ g(SR^{i,j,k}, x, y) \mid \textit{The pixel at } (x,y) \textit{ is an object pixel in } SR^{i,j,k} \} \quad (3)$$

where $g(SR^{i,j,k}, x, y)$ is the gray illumination 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 $[0, M_V - 1]$. 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, denoted by $\mathcal{P}$.

A. Initial Plane Selection Phase

In the first processing phase, to improve the speed and accurateness of the final convergence of the multi-plane region matching and assembling process, the mountain clustering technique [9] can be applied to determine the *SR*s with the most prominent and representative gray intensity features, and these *SR*s are selected as seeds to establish a set of initial planes. The mountain method is a fast, one-pass algorithm, which utilizes the density of features to determine the most representative feature points. Here we consider *SR*s as feature points in the mountain method.

First, given the localized multilevel thresholding process to segment the image into r *SR*s in the *Pool*, the mean $\mu(SR^{i,j,k})$ associated with each of them is also obtained. $\mu(SR^{i,j,k})$ is the mean of gray intensities of object pixels comprised by $SR^{i,j,k}$, and is equivalent to $\mu_k^{i,j}$ obtained by localized multilevel thresholding process. Then the region dissimilarity measure, denoted by D_{RM}, between two 4-adjacent *SR*s can be computed as,

$$D_{RM}(SR^{i_1,j_1,k_1}, SR^{i_2,j_2,k_2}) = \left\| \mu(SR^{i_1,j_1,k_1}) - \mu(SR^{i_2,j_2,k_2}) \right\|, \quad (4)$$

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.

Therefore, the mountain function at a *SR* can be computed as,

$$M_0(SR^{i',j',k'}) = \sum_{\forall SR^{i,j,k} \in Pool} e^{-\alpha D_{RM}(SR^{i,j,k}, SR^{i',j',k'})} \quad (5)$$

where α is a positive constant. A higher value of the mountain function reflects that $SR^{i',j',k'}$ possesses more homogenous *SR*s in its vicinity. Therefore, it is sensible to

select a $SR^{i',j',k'}$ with a high value of mountain function as a representative seed to establish a plane. Let M_0^* be the maximal value of the mountain function values, and SR_0^* be the *SR* whose mountain value is M_0^*:

$$M_0^*\left(SR_0^*\right)=\max_{SR^{i',j',k'}}\left[M_0\left(SR^{i',j',k'}\right)\right] \tag{6}$$

Thus, SR_0^* is selected as the seed of the first initial plane.

After computing the mountain function of each *SR* in the *Pool*, the following representative seeded *SR*s are determined by destroying the mountains. Since the *SR*s whose gray intensity features close to SR_0^* also have high mountain values, it is necessary to eliminate the 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 previous (m-1)th seeded *SR* - SR_{m-1}^*, is computed by,

$$M_m\left(SR^{i',j',k'}\right)=M_{m-1}\left(SR^{i',j',k'}\right)-M_{m-1}^*\left(SR_{m-1}^*\right)e^{-\beta D_{RM}\left(SR^{i,j,k},SR_{m-1}^*\right)} \tag{7}$$

where the parameter β determines the neighborhood radius that provide measurable reductions in the updated mountain function. Then recursively performing the discounted process of the mountain function given by Eq. (7), new suitable seeded *SR*s can be determined in the same way, until the level of the current maximal M_{m-1}^* falls bellow a certain level compared to the first maximal mountain M_0^*. The terminative criterion of this procedure is defined as,

$$\left(M_{m-1}^*/M_0^*\right)<\delta \tag{8}$$

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 [10]. As a result, this process converges to the determination of resultant N seeded *SR*s: $\left\{SR_m^*,\ m=0:N-1\right\}$, and they are utilized to establish N initial planes for performing the following *matching phase*.

B. Matching Phase

In the matching phase, the similarity and connectedness between each current unclassified *SR* and all current exiting planes are analyzed to determine the best belonging plane. If the best matching plane for a certain unclassified *SR* is determined, then this *SR* is assembled into that plane and removed from the *Pool*; otherwise, if there is no matching plane for a certain unclassified *SR*, then this *SR* remains in the *Pool*. Since new planes will be established in the plane constructing phase of the following recursions, the *SR*s which cannot find matching planes in the current recursion of the matching phase will be analyzed in subsequent recursions until they find their best matching plane. Two measurements of the continuity and similarity between two 4-adjacent *SR*s – the side-match measure, denoted as D_{SM}, and the region dissimilarity D_{RM}, as computed by Eq. (4), are employed for the assembling process. Then both D_{SM} and D_{RM} measures are considered to determine the *match grade* of two 4-adjacent *SR*s. The matching phase process is based on evaluating the match grade among the *SR*s and assembling them into planes.

First, the side-match measure - D_{SM}, which reveals the dissimilarity of the touching boundary between the two 4-adjacent *SR*s, is described as follows. Suppose that two *SR*s are 4-adjacent. They may have one of two types of touching boundaries: 1) the vertical touching boundary shared by two horizontally adjacent *SR*s, or 2) the horizontal side shared by two vertically adjacent *SR*s. Here we take the case of two horizontally adjacent *SR*s for example, and the case of vertically adjacent *SR*s can also be similarly derived. For a pair of two horizontally adjacent *SR*s – SR^{i_1,j_1,k_1} on the left, and SR^{i_2,j_2,k_2} on the right, the pixel values on the rightmost side of SR^{i_1,j_1,k_1} and the leftmost side of SR^{i_2,j_2,k_2} can be described as: $g(SR^{i_1,j_1,k_1},M_H-1,y)$ and $g(SR^{i_2,j_2,k_2},0,y)$, respectively. The sets of object pixels on the rightmost side and the leftmost side of an *SR*, denoted by $\boldsymbol{RS}(SR^{i,j,k})$ and $\boldsymbol{LS}(SR^{i,j,k})$, respectively, are defined as follows,

$$\boldsymbol{RS}(SR^{i,j,k})=\left\{g(SR^{i,j,k},M_H-1,y)\middle|\ g(SR^{i,j,k},M_H-1,y)\in\boldsymbol{OP}(SR^{i,j,k}),\text{and }0\le y\le M_V-1\right\} \tag{9}$$

$$\boldsymbol{LS}(SR^{i,j,k})=\left\{g(SR^{i,j,k},0,y)\middle|\ g(SR^{i,j,k},0,y)\in\boldsymbol{OP}(SR^{i,j,k}),\text{ and }0\le y\le M_V-1\right\} \tag{10}$$

To facilitate the following descriptions of the side-match features, the denotations of SR^{i_1,j_1,k_1} and SR^{i_2,j_2,k_2} 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,

$$\boldsymbol{VB}(SR^l,SR^r)=\left\{\left(g(SR^l,M_H-1,y),g(SR^r,0,y)\right)\middle|\ g(SR^l,M_H-1,y)\in\boldsymbol{RS}(SR^l),\text{ and }g(SR^r,0,y)\in\boldsymbol{LS}(SR^r)\right\} \tag{11}$$

Also, the number of side connections of the touching boundary, i.e. the amount of connected pixel pairs in $\boldsymbol{VB}(SR^{i_1,j_1,k_1},SR^{i_2,j_2,k_2})$, should also be considered for the connectedness 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 side-match measure, D_{SM}, of the two 4-adjacent *SR*s can be computed as,

$$D_{SM}(SR^l,SR^r)=\frac{\sum_{\left(g(SR^l,M_H-1,y),g(SR^r,0,y)\right)\in\boldsymbol{VB}(SR^l,SR^r)}\left\|g(SR^l,M_H-1,y)-g(SR^r,0,y)\right\|}{N_{sc}(SR^l,SR^r)} \tag{12}$$

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 should belong to the same object plane $\mathcal{P}$. The range of D_{SM} values is within $[0,255]$.

Accordingly, the D_{SM} measure can reflect the continuity of two 4-adjacent *SR*s, and the D_{RM} value, as obtained by Eq. (4), reveals the similarity between them. Hence the homogeneity and connectedness of two 4-adjacent *SR*s can be measured by determining the dominant effect of the D_{SM}

and the D_{RM}. Therefore, based on the above definitions, the match grade of two 4-adjacent *SR*s, denoted by $\mathfrak{m}$, is determined as,

$$\mathfrak{m}(SR^{i_1,j_1,k_1}, SR^{i_2,j_2,k_2}) = \frac{\max(D_{SM}(SR^{i_1,j_1,k_1}, SR^{i_2,j_2,k_2}),\ D_{RM}(SR^{i_1,j_1,k_1}, SR^{i_2,j_2,k_2}))}{\max(\sigma(SR^{i_1,j_1,k_1}) + \sigma(SR^{i_2,j_2,k_2}),\ 1)} \quad (13)$$

where $\sigma(SR^{i,j,k})$ is the standard deviation of gray intensities of all object pixels associated with the $SR^{i,j,k}$, and is equivalent to $\sigma_k^{i,j}$ obtained from localized histogram multilevel thresholding process. Here the denominator term $\max(\sigma(SR^{i_1,j_1,k_1}) + \sigma(SR^{i_2,j_2,k_2}),\ 1)$ in Eq. (13) serves as the normalization factor.

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 match grades of $SR^{i,j,k}$ associated with those *SR*s, denoted by $SR_q^{i',j',k'}$, which have been grouped into current existing object planes (the subscript q represents that $SR_q^{i',j',k'}$ belongs to the q-th plane $\mathcal{P}_q$) to seek for the matching plane into which $SR^{i,j,k}$ can be grouped. In order to facilitate match grade determination, the set $\boldsymbol{AS}(SR^{i,j,k}, \mathcal{P}_q)$ is utilized for containing the $SR_q s$ in a plane $\mathcal{P}_q$ which are 4-adjacent to $SR^{i,j,k}$, is defined by,

$$\boldsymbol{AS}(SR^{i,j,k}, \mathcal{P}_q) = \left\{ SR_q^{i',j',k'} \in \mathcal{P}_q \,\middle|\, SR_q^{i',j',k'} \text{ is } 4-adjacent \text{ to } SR^{i,j,k} \right\} \quad (14)$$

Then the match grade $\mathfrak{M}(SR^{i,j,k}, \mathcal{P}_q)$, which reveals how well $SR^{i,j,k}$ matches with $\mathcal{P}_q$, can be determined by the following operation,

$$\mathfrak{M}(SR^{i,j,k}, \mathcal{P}_q) = \min_{\forall SR_q^{i',j',k'} \in \boldsymbol{AS}(SR^{i,j,k}, \mathcal{P}_q)} \mathfrak{m}(SR^{i,j,k}, SR_q^{i',j',k'}) \quad (15)$$

It must be noted that if none of the $SR_q s$ in $\mathcal{P}_q$ are 4-adjacent to $SR^{i,j,k}$, i.e. $\boldsymbol{AS}(SR^{i,j,k}, \mathcal{P}_q) = \phi$, then $\mathcal{P}_q$ is excluded from the consideration for matching with $SR^{i,j,k}$. Thus the plane which has the best match grade associated with $SR^{i,j,k}$ among all existing planes, denoted by $\mathcal{P}_{\mathrm{m}}$, can be determined by,

$$\mathfrak{M}(SR^{i,j,k}, \mathcal{P}_{\mathrm{m}}) = \min_{\forall \mathcal{P}_q} \mathfrak{M}(SR^{i,j,k}, \mathcal{P}_q) \quad (16)$$

The following *matching condition* is then applied to check whether the selected candidate plane $\mathcal{P}_{\mathrm{m}}$ and $SR^{i,j,k}$ are sufficiently matched, and thus $\mathcal{P}_{\mathrm{m}}$ can be determined for suitably absorbing $SR^{i,j,k}$, defined as follows,

$$\mathfrak{M}(SR^{i,j,k}, \mathcal{P}_{\mathrm{m}}) \le Th_{\mathrm{m}} \quad (17)$$

where Th_{m} is a predefined threshold which represents the acceptable tolerance of dissimilarity for $SR^{i,j,k}$ to be grouped into $\mathcal{P}_{\mathrm{m}}$. The matching condition has a moderate effect on the number of resultant object planes, and the value choice of $Th_{\mathrm{m}} = 1.2$ is experimentally determined to obtain sufficiently distinct planes and avoid excessive splitting of planes. Accordingly, if $SR^{i,j,k}$ and its associated $\mathcal{P}_{\mathrm{m}}$ satisfy the matching condition, then $SR^{i,j,k}$ is merged into $\mathcal{P}_{\mathrm{m}}$, and removed from the *Pool*. If the matching condition cannot be satisfied, this reflects that there is no appropriate matching plane for $SR^{i,j,k}$ in this current matching phase recursion. As a result, $SR^{i,j,k}$ should remain in the *Pool* until its suitable matching plane emerges after more plane constructing phase recursions. After a 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 have been processed once in the current matching phase recursion.

C. Plane Constructing Decision Phase

After performing the previous matching phase recursion, if there are unclassified *SR*s remaining, and the *Pool* is not drained, these unclassified *SR*s must be analyzed and a determination made as to whether it is necessary to establish a new plane to assemble *SR*s with such features into another homogeneous region. The plane constructing decision phase determines whether to 1) establish and initialize a new plane by selecting the unclassified *SR* "farthest" from all existing planes as an initial seed, or 2) extend one selected plane by merging one unclassified *SR* "nearest" to this plane. The decision is made according to the analysis of the following gray intensity and location features. The dissimilarity measure between one unclassified *SR*, not adjoined to any existing planes in the previous matching phase recursion, and a certain object plane $\mathcal{P}_q$, is determined by the relative-difference of gray intensities between this *SR* and its nearest $SR_q^{i',j',k'}$ among all $SR_q s$ that belong to $\mathcal{P}_q$, and is computed as,

$$D_I(SR^{i,j,k}, \mathcal{P}_q) = \min_{\forall SR_q \in \mathcal{P}_q} \left(\frac{D_{RM}(SR^{i,j,k}, SR_q^{i',j',k'})}{\max(\sigma(SR^{i,j,k}) + \sigma(SR_q^{i',j',k'}),\ 1)} \right) \quad (18)$$

where $D_{RM}(SR^{i,j,k}, SR_q^{i',j',k'})$ is computed by Eq. (4). Then the smallest dissimilarity of gray intensity between $SR^{i,j,k}$ and the plane being most similar to it among all currently existing object planes, denoted by $\mathcal{P}_{SI(SR^{i,j,k})}$, is determined by,

$$D_I(SR^{i,j,k}, \mathcal{P}_{SI(SR^{i,j,k})}) = \min_{\forall \mathcal{P}_q}(D_I(SR^{i,j,k}, \mathcal{P}_q)), \quad (19)$$

Then the dissimilarity measure of $SR^{i,j,k}$ and its associated $\mathcal{P}_{SI(SR^{i,j,k})}$, is also expressed as $D_{SI}(SR^{i,j,k})$, to represent the least dissimilarity of gray intensity between this *SR* and all existing planes. If $SR^{i,j,k}$ has a sufficiently low dissimilarity D_I with the plane $\mathcal{P}_{SI(SR^{i,j,k})}$, and they are also locatively close to each other, this means that this *SR* is homogeneous with $\mathcal{P}_{SI(SR^{i,j,k})}$, even if it is not currently 4-adjacent to $\mathcal{P}_{SI(SR^{i,j,k})}$.

To determine this situation, the locative distance between an *SR* and a plane $\mathcal{P}_q$, denoted as $D_E(SR^{i,j,k}, \mathcal{P}_q)$, is computed by the Euclidean distance between this *SR* and the

closest SR_q among all $SR_q s$ associated with the plane $\mathcal{P}_q$; and is determined as,

$$D_E(SR^{i,j,k}, \mathcal{P}_q) = \min_{\forall SR_q \in \mathcal{P}_q} D_e(SR^{i,j,k}, SR_q^{i',j',k'}), \quad (20)$$

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

If $SR^{i,j,k}$ and its $\mathcal{P}_{SI(SR^{i,j,k})}$ are homogeneous in gray illumination and also locatively close to each other, i.e. both $D_I(SR^{i,j,k}, \mathcal{P}_{SI(SR^{i,j,k})})$ and $D_E(SR^{i,j,k}, \mathcal{P}_{SI(SR^{i,j,k})})$ values are sufficiently small, then $SR^{i,j,k}$ should join the plane $\mathcal{P}_{SI(SR^{i,j,k})}$, rather than establish a new independent plane, in order to prevent a text region or homogeneous object to be split into more than one plane. Otherwise, if no such planes are found, a new plane should be created to aggregate those *SR*s with distinct features.

To ensure that the new plane contains distinct features with currently existing planes, a scheme for selecting the suitable *SR* as the representative seed for constructing a new plane is given as follows. By this means, this seed *SR*, which is most dissimilar to any currently existing planes in illumination intensities, can be obtained by,

$$D_{LI}(SR^{LI}) = \max_{\forall SR \in Pool} D_{LI}(SR^{i,j,k}), \quad (21)$$

where $D_{LI}(SR^{i,j,k}) = \max_{\forall \mathcal{P}_q}(D_I(SR^{i,j,k}, \mathcal{P}_q))$

Hence, the determined SR^{LI} with the largest D_{LI} value will be selected as the seed *SR* to establish a new plane to aggregate those *SR*s whose features are distinct from other existing planes.

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

Step 1: First, the *SR*s which have sufficiently low D_{SI} values are selected into the set $\boldsymbol{SR_{SI}}$ using the following operation:

$$\boldsymbol{SR_{SI}} = \left\{ SR^{i,j,k} \in Pool \,\middle|\, D_{SI}(SR^{i,j,k}) \le Th_{SI} \right\}, \quad (22)$$

where Th_{SI} is a predefined threshold for determining whether the *SR* is homogeneous with any one of existing planes. If none of the *SR*s are selected by the above condition, i.e. $\boldsymbol{SR_{SI}}$ is empty, then go directly to **Step 3** for constructing a new plane; otherwise, perform the following **Step 2**.

Step 2: The set $\boldsymbol{SR_{SI}}$ now contains the *SR*s which are homogeneous with other currently existing planes, but are not 4-adjacent with them, and thus remain unclassified in the previous matching phase recursion. The *SR* locatively nearest to its associated $\mathcal{P}_{SI(SR^{i,j,k})}$, denoted by SR^N, is determined as follows,

$$D_E(SR^N, \mathcal{P}_{SI(SR^N)}) = \min_{\forall SR \in \boldsymbol{SR_{SI}}} D_E(SR^{i,j,k}, \mathcal{P}_{SI(SR^{i,j,k})}) \quad (23)$$

If SR^N and its $\mathcal{P}_{SI(SR^N)}$ are sufficiently close to each other, i.e. the condition $D_E(SR^N, \mathcal{P}_{SI(SR^N)}) \le Th_L$ is satisfied, then SR^N is decided to be merged with $\mathcal{P}_{SI(SR^N)}$ to extend its influential area on nearby *SR*s, and proceeds to **Step 4**. Otherwise, perform **Step 3** for constructing a new plane.

Step 3: The SR^{LI}, the *SR* most dissimilar to any currently existing planes, is determined by Eq. (21). Thus SR^{LI} is employed as a seeded *SR* to establish a new plane $\mathcal{P}_{new}$, and then continues to **Step 4**.

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

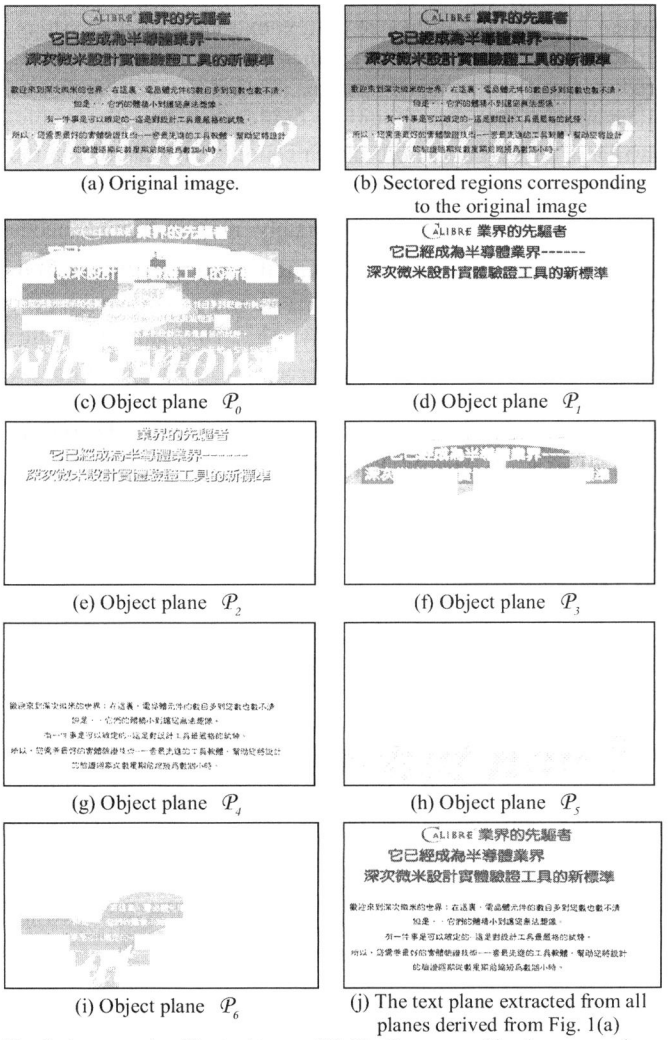
The threshold Th_{SI} utilized in Eq. (22) moderately influences the number of resultant planes. If Th_{SI} is low, then the number of resultant planes will be increased and a homogeneous region may be broken into more than one plane, although its influence on text extraction is not serious. If the value is large however, then the number of planes is reduced, and some objects may be merged to a certain degree. Reasonably, its value should be tighter than the Th_m value, which is utilized in the pre-match condition to ensure that the determined SR^N is sufficiently homogeneous with $\mathcal{P}_{SI(SR^N)}$ and thus $\mathcal{P}_{SI(SR^N)}$ can appropriately absorb homogeneous *SR*s near the extended influential area benefited from participation with SR^N. Therefore, in our experiments, the value of Th_{SI} is chosen as $(3/4) \cdot Th_m$. Normally, text-lines or text-blocks usually occupy perceptible area of the image, and thus their width or height should be in appreciable proportion to those of the whole image. Therefore, $Th_L = \min(N_H, N_V)/4$ is used for experiments, where N_H and N_V are the numbers of block regions per row and per column, respectively.

D. Overall Process

Based on the three above-mentioned processing phases, the region matching and assembling process begins by applying the initial plane selection phase on all unclassified *SR*s in the *Pool* to determine the representative seeded *SR*s $\{SR_m^*, m = 1:N\}$ for setting up N initial planes. Thus, the matching phase can then be carried out for the rest of *SR*s in the *Pool* and the initial planes $\mathcal{P}_0$, ..., $\mathcal{P}_{N-1}$. Then the matching phase and the plane constructing decision phase are recursively performed in turns on the rest of the *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.

Consequently, after the multi-plane region matching and assembling process is completed, each of the homogenous objects with connected and similar features is separated into corresponding object planes. To facilitate further analysis, they are represented as $\mathcal{P}_0$, $\mathcal{P}_1$, ..., $\mathcal{P}_{L-1}$, where L is the number of the resultant planes obtained. We use Fig. 1 as an example of the proposed method procedure. The original image in Fig. 1(a) consists of three different colored text regions printed on a varying and shaded background. Moreover, the black characters are superimposed on the white characters. First, as shown in Fig. 1(b), the original image is transformed into grayscale, and is divided into block regions. Figures 1(c)-(i) are seven object planes obtained from Fig. 1(a) by performing the multi-plane segmentation technique. As a result, the homogeneous objects in which all characters

and background textures are segmented into several separate planes can effectively be analyzed in detail. By observing these obtained planes, we can see that three text regions with different characteristics are distinctly separated. Extraction of text strings from each binarized plane in which objects are well separated, and can be easily performed by our previous proposed projection-based text extraction method [11]. After the text extraction process being conducted on all object planes, the text lines extracted from these planes are then collected into a resultant text plane, as shown in Fig. 1(j).



(a) Original image. (b) Sectored regions corresponding to the original image

(c) Object plane $\mathcal{P}_0$ (d) Object plane $\mathcal{P}_1$

(e) Object plane $\mathcal{P}_2$ (f) Object plane $\mathcal{P}_3$

(g) Object plane $\mathcal{P}_4$ (h) Object plane $\mathcal{P}_5$

(i) Object plane $\mathcal{P}_6$ (j) The text plane extracted from all planes derived from Fig. 1(a)

Fig. 1. An example of the test image, "Calibre", processed by the proposed multi-plane segmentation technique (image size: 1929 x 1019)

IV. Experimental Results

The performance of the proposed method is evaluated in this section and compared to other existing text extraction methods, namely Jain and Yu's color-quantization-based method [6], and Pietikainen and Okun's edge-based method [4]. A set of 46 real-life complex document images was employed for experiments on performance evaluation of text extraction [12]. These test images are scanned from book covers, book and magazine pages, advertisements, and other real-life documents at the resolution of 200 dpi to 300 dpi. They were transformed into gray-scale images, and then processed by the proposed method. Most are comprised of character strings in various colors or illuminations, font styles and sizes which are overlapped with multi-colored, textured, or uneven illuminated backgrounds. Figures 2(a) – 3(a) are the most representative samples with typical characteristics of complex document images, and more results of test samples in the experimental set are also shown in [12].

Figures 2 – 3 show the results of text extraction produced by the proposed method, Jain and Yu's color-quantization-based method [6], and Pietikainen and Okun's edge-based method [4]. Here the extraction results of Pietikainen and Okun's method in Figs 2(d) – 3(d) were converted into masked images where the black mask was adopted to exhibit the non-text region. Figure 2(a) contains background objects with sharp illumination variations across text regions, and some of these also possess similar colors and illuminations to those characters touched with them. As a result, the character illuminations are influenced and have gradational variations due to the scanning process. After performing the proposed method, the extraction results shown in Fig. 2(b) demonstrate that the majority of the characters are successfully segmented from the sharply varying backgrounds. As shown in Fig. 2(c), Jain and Yu's method fails to extract the large captions and many characters of the main text region, because many characters are fragmented due to the influence of those background objects in color-quantization process. The edge-based method extracts most characters except some broken large characters and several missed small characters, as shown in Fig. 2(d); however, several graphical objects with sharply varying contours are also identified as text, and thus the characters in extracted text regions are blurred. In Fig. 3(a), large portions of the main body texts are printed on a large black textured and shaded region, and thus the contrast between the characters and this textured region is extremely degraded. As shown in Fig. 3(b), the characters in three different text regions are successfully extracted by the proposed method. Both Jain and Yu's method and the edge-based method fail to extract characters from the textured region with degraded contrast, as shown in Figs. 3(c) and 3(d), respectively. Therefore, as seen from Figs. 2 – 3, our proposed method performs significantly better than Jain and Yu's method and the edge-based method in various difficult cases. By observing the results obtained, while character strings comprised of various illuminations, sizes, and styles, are overlapped with background objects with considerable variations in contrast, illumination, and texture, nearly all the text regions are effectively extracted by the proposed method.

To quantitatively evaluate text extraction performance, two measures, the recall rate and the precision rate, which are commonly used for evaluating performance in information retrieval, are adopted. They are respectively defined as,

$$recall\ \ rate = \frac{No.\ of\ \ correctly\ \ extracted\ \ characters}{No.\ of\ \ actual\ \ characters} \tag{24}$$

$$precision\ \ rate = \frac{No.\ of\ \ correctly\ \ extracted\ \ characters}{No.\ of\ \ extracted\ \ character-like\ components} \tag{25}$$

We compute the recall and precision rates for text extraction results of test images in this study by manually counting the number of actual characters of the document image, total

extracted character-like components, and the correctly extracted characters, respectively. The quantitative evaluation were performed on our test set [12] of 46 complex document images totaling 22791 visible characters. Since these quantitative evaluation criteria are performed on the extracted connected-components, the results of Pietikainen and Okun's method [4] is inappropriate for quantitative evaluation using these criteria. Table 1 depicts the results of quantitative evaluation of Jain and Yu's method [6] and the proposed method. By observing Table 1, we can see that the proposed method exhibits significantly better text extraction performance as compared with that of Jain and Yu's method.

Table I. Experimental data of Jain and Yu's method and proposed method

Method	Recall Rate	Precision Rate
Jain and Yu's method	79.8%	95.2%
Proposed method	99.1 %	99.3%

REFERENCES

[1] L. O' Gorman, R. Kasturi, Document image analysis, IEEE Comput. Soc. Press, Silver Spring, MD, 1995.

[2] V. Wu, R. Manmatha, and E.M. Riseman, "Textfinder: an automatic system to detect and recognize text in images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 11, pp. 1224-1229, 1999.

[3] Y. M. Y. Hasan, and L. J. Karam, "Morphological Text Extraction from Images," *IEEE Trans. Image Process.*, vol. 9, no. 11, pp. 1978-1983, 2000.

[4] M. Pietikinen and O. Okun, "Edge-based method for text detection from complex document images," *Proc. 6th Int'l Conf. Doc. Anal. Recognit.*, pp. 286-291, 2001.

[5] Y. Zhong, K. Karu, A. K. Jain, "Locating text in complex color images," *Pattern Recognit.*, vol. 28, no. 10, pp. 1523-1535, 1995.

[6] A. K. Jain, B. Yu, "Automatic text location in images and video frames," *Pattern Recognit.*, vol. 31, no. 12, pp. 2055-2076, 1998.

[7] C. Strouthopoulos, N. Papamarkos, A. E. Atsalakis, "Text extraction in complex color documents," *Pattern Recognit.*, vol. 35, pp. 1743-1758, 2002.

[8] B.-F. Wu, Y.-L. Chen, and C.-C. Chiu, "A discriminant analysis based recursive automatic thresholding approach for image segmentation," *IEICE Trans. Info. Systems*, vol. E88-D, no.7, pp.1716-1723, 2005.

[9] R.R. Yager and D.P. Filev, "Approximate clustering via the mountain method," *IEEE Trans. Syst. Man Cybern*, vol. 24, no. 8, pp. 1279-1284, 1994.

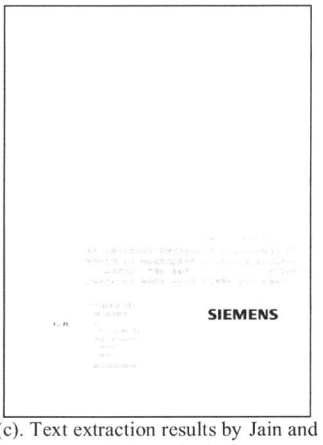
[10] N.R. Pal and D. Chakraborty, "Mountain and subtractive clustering method: improvements and generalization," *Int'l. J. Intell. Syst.*, vol. 15, pp. 329-341, 2000.

[11] B.-F. Wu, Y.-L. Chen, and C.-C. Chiu, "Multi-layer segmentation method for complex document images," *Int'l. J.* Pattern Recognit. Artif. Intell., Vol. 19, No. 8, pp. 997-1025, 2005.

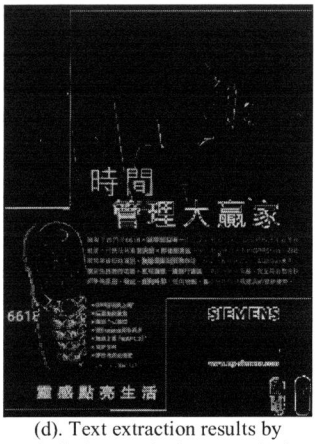
[12] Experimental results of all test images in our database are available at: http://140.113.150.97/SMC06_TestDatabase.htm

(a). Original Image

(b). Text extraction results by the proposed method

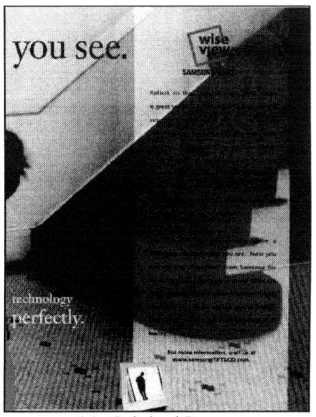


(c). Text extraction results by Jain and Yu's method



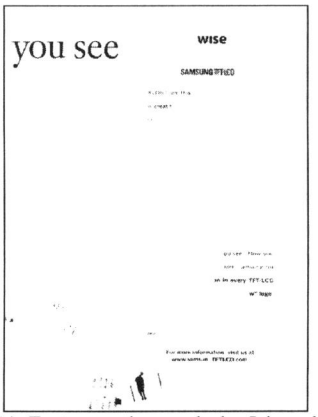
(d). Text extraction results by Pietikainen and Okun's method

Fig. 2. Results of the test image 1 (size: 2333 × 3153)

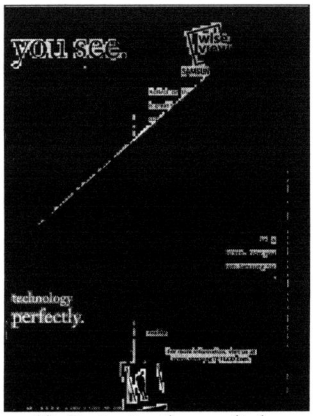


(a). Original Image

(b). Text extraction results by the proposed method



(c). Text extraction results by Jain and Yu's method



(d). Text extraction results by Pietikainen and Okun's method

Fig. 3. Results of the test image 2 (size: 2405 × 3207)