Text, photo, and line extraction in scanned documents

Article *in* Journal of Electronic Imaging · July 2012

DOI: 10.1117/1.JEI.21.3.033006

CITATIONS
5

READS
335

5 authors, including:

Sezer Erkilinc
University College London
59 PUBLICATIONS **460** CITATIONS
SEE PROFILE

Mustafa Jaber
Rochester Institute of Technology
19 PUBLICATIONS **38** CITATIONS
SEE PROFILE

Eli Saber
Rochester Institute of Technology
125 PUBLICATIONS **3,377** CITATIONS
SEE PROFILE

Journal of Electronic Imaging

SPIEDigitalLibrary.org/jei

Text, photo, and line extraction in scanned documents

M. Sezer Erkilinc
Mustafa Jaber
Eli Saber
Peter Bauer
Dejan Depalov

Text, photo, and line extraction in scanned documents

M. Sezer Erkilinc
University College London
Department of Electronic and Electrical Engineering
Optical Networks Group
Torrington Place
London WC1E 7JE, United Kingdom
E-mail: m.erkilinc@ee.ucl.ac.uk

Mustafa Jaber
IPPLEX Holdings Corporation
Santa Monica
California 90025

Eli Saber
Rochester Institute of Technology
Department of Electrical and Microelectronic Engineering
Rochester, New York 14623

Peter Bauer
Dejan Depalov
Hewlett-Packard Corporation
Imaging Asset Team
Boise, Idaho 83714

Abstract. *We propose a page layout analysis algorithm to classify a scanned document into different regions such as text, photo, or strong lines. The proposed scheme consists of five modules. The first module performs several image preprocessing techniques such as image scaling, filtering, color space conversion, and gamma correction to enhance the scanned image quality and reduce the computation time in later stages. Text detection is applied in the second module wherein wavelet transform and run-length encoding are employed to generate and validate text regions, respectively. The third module uses a Markov random field based block-wise segmentation that employs a basis vector projection technique with maximum* a posteriori *probability optimization to detect photo regions. In the fourth module, methods for edge detection, edge linking, line-segment fitting, and Hough transform are utilized to detect strong edges and lines. In the last module, the resultant text, photo, and edge maps are combined to generate a page layout map using K-Means clustering. The proposed algorithm has been tested on several hundred documents that contain simple and complex page layout structures and contents such as articles, magazines, business cards, dictionaries, and newsletters, and compared against state-of-the-art page-segmentation techniques with benchmark performance. The results indicate that our methodology achieves an average of ~89% classification accuracy in text, photo, and background regions.* © *2012 SPIE and IS&T.* [DOI: 10.1117/1.JEI.21.3.033006]

Paper 11318 received Nov. 23, 2011; revised manuscript received May 15, 2012; accepted for publication Jun. 8, 2012; published online Jul. 13, 2012.

1 Introduction

Document image understanding and page layout analysis have become an important application since the beginning of the 1990s[1] as document databases started to shift from hard- to soft-copy with the arrival of fast computers, large memory chips, and inexpensive scanners. Methodologies, algorithms, and systems are being developed to mine information from document images by locating and extracting line, photo, and/or text regions hierarchically in a "human-like" fashion. The field of document image understanding covers a variety of document types such as technical articles, manuals, outlines, mail-pieces, drawings, and maps to name a few. Page layout classification systems are being utilized in many applications such as object-oriented rendering[2] in order to enable efficient cartridge usage while printing documents at different resolutions. Other applications include document retrieval,[3–5] where effective memory usage and quick access are essential for practical solutions. Optical character recognition (OCR)[2–8] is yet another application where handwritten, typewritten, or printed text is converted into machine-encoded text for further processing.

Page classification is one of the main explored topics in document processing to drive homogeneity criteria for connected regions of text, graphic/photo, and background. It is generally used as an initial step for page layout analysis, OCR and/or document retrieval. There are three main approaches in page segmentation, namely top-down, bottom-up, and hybrid approaches. Generally speaking, a top-down approach analyzes the global information found on

the page for the purpose of splitting the document into blocks, blocks into lines, and lines into words. To this effect, techniques for classifying documents into smaller components involve obtaining statistics by run-length encoding (RLE)[9] and extracting geometrical characteristics by utilizing page layout features.[10–12] Alternative schemes include finding rectangular blocks and applying vertical and horizontal projections to a document image using local connectivity property.[13,14]

Conversely, a bottom-up approach utilizes the local information such as connected components in a specific region or block to perform the classification. It first locates individual words, then merges words into lines, lines into blocks, and finally blocks into columns. Labeled rectangular regions are then extracted using meaningful features and are subsequently connected by applying adaptive classification schemes.[15,16] However, the main drawback of Refs. 15 and 16 is that segmentation is achieved based on the assumption that a document image consists of rectangular areas. To overcome this restriction, the structure of white-spaced background, surrounded by printed zones, is identified and utilized to segment the document.[17–19] In Ref. 20, a special distance-metric between the page components (text and photo regions) is used to construct a physical page structure. In addition, the technique in Ref. 21 employes sharp edges as features to separate textual and non-textual regions.

Hybrid approaches, on the other hand, achieve page segmentation and classification based on features extracted directly from the document. Initially, the document image is subdivided into blocks that are, in turn, utilized to compute appropriate coefficients or features. Applying two-dimensional (2-D) Gabor filter on subdivided blocks to obtain coefficients,[22] computing geometric and texture characteristics,[23] finding local subband energy features,[24] extracting features from a connected component histogram,[25] and utilizing gray level cooccurrence matrix (GLCM)[26] are some of the hybrid approaches to segment and classify the document image. In Refs. 27 and 28, a larger set of features for page layout classification was proposed where a forward selection method is iteratively employed to compute a linear feature until the desired accuracy is achieved. This automatic feature extraction is used to analyze several document databases that contain images, graphics, handwriting, and machine-printed text regions.

All the studies mentioned above address page or document segmentation/classification problem in an unsupervised way. In other words, they do not require any *a priori* information. Unsupervised segmentation was employed particularly at the end of the 1980s and the beginning of the 1990s in order to sidestep the computational burden imposed by training since the training phase, unsurprisingly, was a very time consuming stage with 1990's processor technology. However, this computational complexity issue began to subside with 2000's processor technology paving the way for supervised segmentation schemes. Techniques such as hierarchical conditional random fields (CRFs),[29] K-nearest neighbor (K-NN),[30] Fisher classifiers,[31,32] decision-tree classifiers, self-organizing maps,[33] and neuro-fuzzy approaches[34] were subsequently employed in the training stages for page segmentation.

In this paper, we propose a robust and efficient multimodule unsupervised page classification system to detect text, photo, and strong edge/line regions in both complex color and gray-scale scanned documents. The first module encompasses a preprocessing step that includes image down-sampling, color space conversion, gamma correction, and morphological operation. Down-sampling is applied to reduce the computational burden. Moreover, morphological dilation is performed on the lightness (L*) component of CIEL*a*b* color space to enhance text, photo, and strong edge/line detection. Three different modules are then employed in parallel to process the L* component of the scanned document to detect text, photo, and strong edge/line regions separately. The photo-detection stage relies on a Markov random field (MRF) block-wise segmentation process, introduced by Won,[35] using a maximum *a posteriori* (MAP) optimization framework. Multilevel wavelet decomposition[36] is utilized to extract features and classify text regions. For verification purposes, RLE, first introduced by Wong et al.[37] is applied to obtain a final text map. Strong edge/line regions are extracted by employing the standard Hough transform on the L* channel. The resultant text, photo, and edge detection maps are then combined using K-Means clustering into a final classification map.

Our algorithm differs from the state-of-the-art page classification methods by requiring no restrictions on the color of background of the test documents, while other published techniques assume that the background is white when extracting text[23] and/or photo[35] regions. Furthermore, our system is not limited to documents with simple background such as textbook pages, newsletters, or handouts. Instead, it is developed to handle several types of scanned documents such as magazines, articles, advertisements, and correspondences that contain complex text, photo, and color background zones. Additionally, our algorithm does not depend on a given scanning technique or resolution. It is designed to cope with documents scanned on both RGB or gray-scale scanners. We also differ from prior art in not only classifying both text and photo regions but also detecting the strong edge and line borders between the regions.

The remainder of the paper is organized as follows: the proposed scheme, with the designated modules for preprocessing, text, photo, strong edge/line detection, and map combination, is outlined in Sec. 2. Results, performance evaluation, and comparison with state-of-the-art methods are presented in Sec. 3. Finally, conclusions are drawn in Sec. 4.

2 Proposed Algorithm

A page layout classification methodology that can process RGB or gray-scale images has been proposed. The algorithm starts by a preprocessing module that includes image enhancement techniques such as filtering, image re-sizing, color space transformation, morphological dilation, and/or gamma correction. To this effect, RGB to CIEL*a*b* color space conversion is utilized for colored input scanned documents, while gamma correction is employed for gray-scale images. Next, the preprocessed lightness channel (L*) is used by the text, photo, and line/strong edge detection modules to generate three segmentation maps. Finally, K-Means clustering is utilized to combine the segmentation maps into a single page layout classification map. A flowchart of the proposed approach is shown in Fig. 1.

2.1 *Preprocessing Module*

This module contains different stages for low-pass filtering, image re-scaling, morphological dilation, and color space

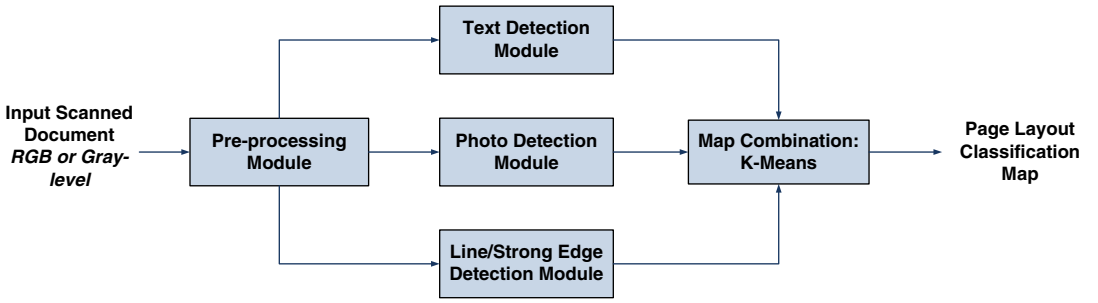


Fig. 1 Flowchart of the proposed approach.

transformation or gamma correction based on the image type. The objectives of the preprocessing module are to prevent aliasing due to down-sampling, reduce computation time, and eliminate noise and illumination variations in the document. A block diagram of the preprocessing module is given in Fig. 2.

In this study, a typical document size of 8.5×11 inch is used. It is scanned at 300 dots per inch (dpi) yielding an input image of size $\sim 3300 \times 2600$ pixels. This image size is empirically chosen to strike the appropriate balance between quality and computational complexity. To reduce the computational burden, the image is low-pass filtered using 11×11 pixels kernel to minimize aliasing artifacts and then down-sampled by a scale factor, $k = 1/4$. The down-sampled image (for color images only) is then converted from RGB to CIEL*a*b*. The main benefits of this transformation are to provide approximate perceptual uniformity and to perform an inherent gamma correction. Therefore, to simulate similar behavior for gray-scale scanned documents, a similar correction is applied to minimize illumination variations and suppress noise in background regions. Its aim is to create a realistic image in terms of shading, intensity, luminance, and/or brightness. A gamma factor (γ) of 2.2, chosen empirically based on our extensive testing, is used in our method.

Finally, a dilation morphological operation is employed to enhance high-frequency regions in the preprocessed L* component of the image. The dilation operation scans the input intensity image to find local maxima in a given direction over a small window. It is applied twice to emphasize high-frequency regions in both horizontal and vertical directions. These two maps ($\text{Dilation}_{\text{Horizontal}}$ and $\text{Dilation}_{\text{Vertical}}$) are averaged and subtracted from the input L* channel as shown in Eq. (1). The $|.|$ sign stands for the absolute value in Eq. (1).

$$\text{Preprocessed L}^*\text{or Gray-scale Image} = \left| L^* - \frac{1}{2}\left(\text{Dilation}_{\text{Horizontal}} + \text{Dilation}_{\text{Vertical}}\right) \right|. \quad (1)$$

2.2 *Text Detection Module*

The text detection module uses the L* channel obtained from the preprocessing stage to classify text regions. First, a multilevel wavelet decomposition is employed to compute the local energy, using a variable window size, in the LH, HL, and HH subbands (horizontal, vertical, and diagonal detail channels). The energy maps are up-sampled to the size of the input preprocessed L* channel and averaged to generate a text-candidate map. As a second step, a RLE module is applied to verify and validate text regions. A block diagram of the proposed module is presented in Fig. 3 and each of the submodules are detailed in the following subsections.

2.2.1 *Multilevel wavelet decomposition and local energy computation*

The goal of this operation is to identify candidate text regions. We employ a basic assumption that text regions tend to display high variations in small neighborhood areas and possess distinct contrast with respect to the background. Our algorithm also handles more challenging cases with complex backgrounds as shown in Sec. 3.

Our technique (see Fig. 4) utilizes the discrete wavelet transform (DWT) to decompose the preprocessed L* channel generated using Eq. (1). Next, the energy is computed using a variable-size sliding window. The window size varies with respect to the original document spatial size and the wavelets decomposition level (8×8 pixels in the first level and

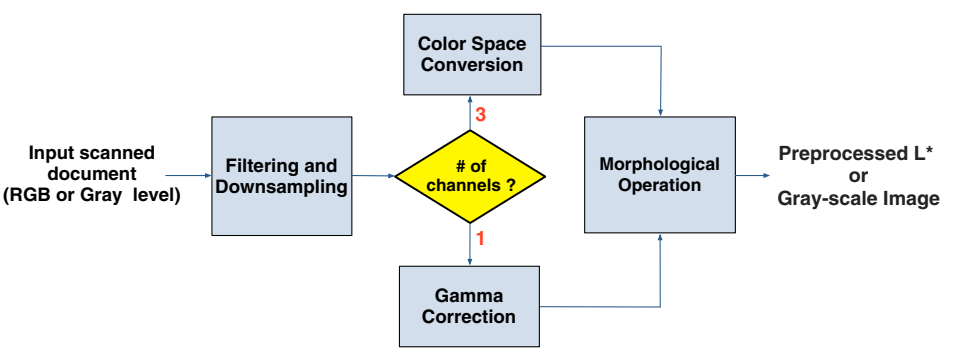


Fig. 2 Block diagram of the preprocessing module.

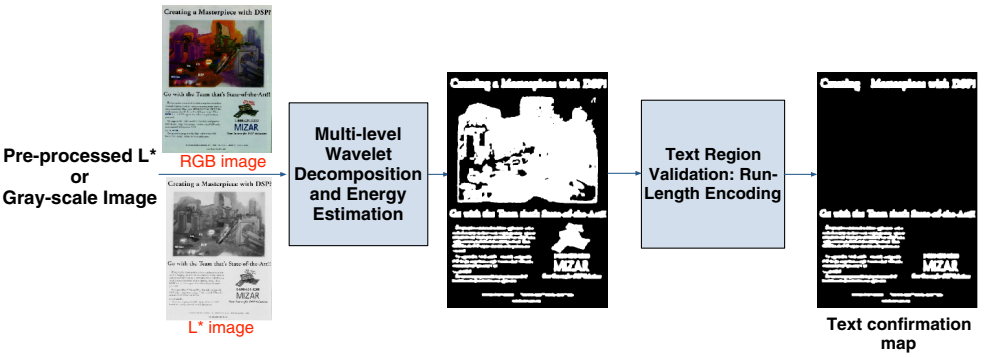


Fig. 3 Block diagram of the text-detection module.

4×4 pixels in the second wavelet decomposition level). Additionally, the average value of the wavelets coefficients in the given window is subtracted from all coefficients to eliminate any bias in L* values toward a certain intensity value. This later step emphasizes the contrast between the text and background regions and prevents the color of background regions from dominating the energy values. The aforementioned procedure is applied twice where the local average of the neighboring coefficients (I'_{local}) and the global average of all coefficients (I'_{global}) in the subband are subtracted from the second term of Eq. (2).

$$\text{Text}_{\text{Energy Map}} = \sqrt{\sum_{x,y \subset W} [I(x,y) - I'_{\text{local}}]^2} + \sqrt{\sum_{x,y \subset W} [I(x,y) - I'_{\text{global}}]^2}, \quad (2)$$

where W stands for the local window and $I(x,y)$ is the wavelet coefficient at the location x and y.

Figure 4 provides a block diagram for the wavelet decomposition and energy map submodule, where two DWT levels using Daubechies 4-tap filter-banks are shown. The range of the energy maps is first normalized before up-sampling these maps to the original document size and the resultant up-sampled maps are averaged. A bi-cubic interpolation (see Sec. 2.1) technique is used to resize the energy maps with a scaling factor (2^s), where s is the wavelet scale (level).

The purpose of the wavelets and energy computation is to generate a text-candidate map that outlines the exact text-candidate regions (binary map). However, these operations generate gray-scale maps that are sensitive to high-frequency regions such as texts, textures, and edges. Therefore, a thresholding operation based on Otsu's method[38] is employed. Some of the target scanned documents in our test database have text printed zones using different colors/gray-levels. Consequently, a histogram equalization operation is applied before the thresholding stage to reduce the energy variation

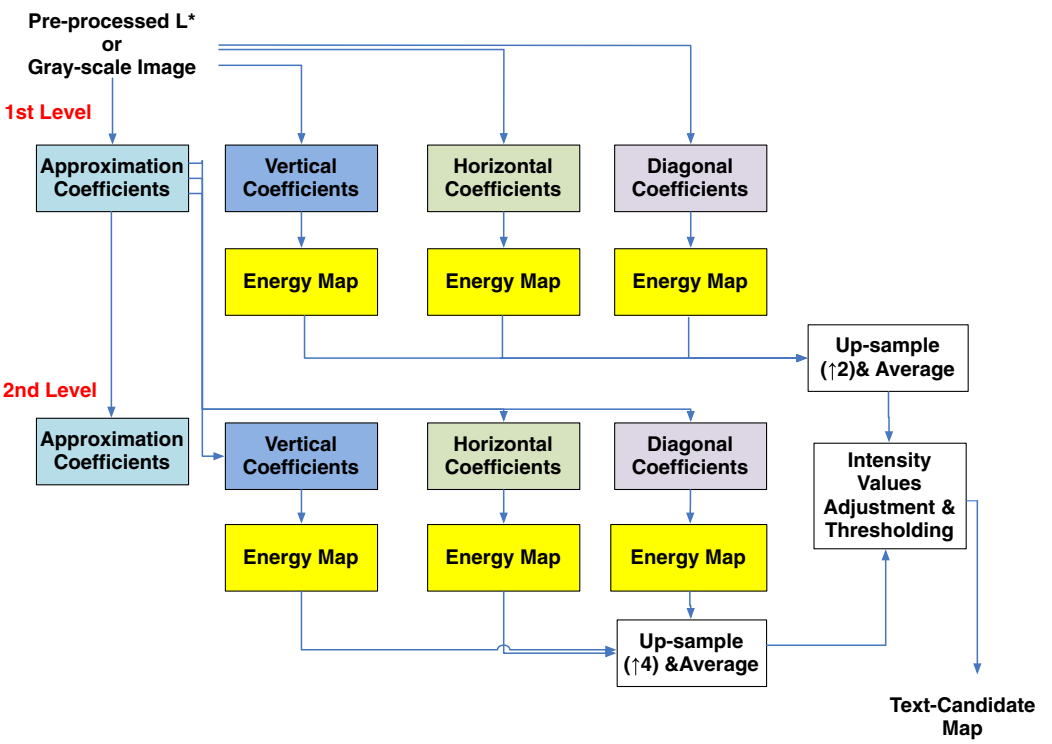


Fig. 4 Block diagram of the multilevel wavelet decomposition and energy map submodule.

due to text color. Another operation, applied to the binary text-candidate regions as a postprocessing step, is the elimination of relatively small regions. In this sense, regions with area less than 0.03% of the original scanned document size are removed from the binary text-candidate map because these small regions do not contribute any significant information.

2.2.2 *Text region validation*

This module uses the text-candidate maps generated in Sec. 2.2.1 and the L* channel of the original scanned document. It is assumed that a text region consists of a sentence, multiple sentences, or a paragraph. Hence, if any text-candidate region is considered by itself, its structure should follow the above-stated assumption and generate a set of peaks and valleys of intensity values if averaged in the horizontal or vertical direction (profile projection). The characteristics of these peaks and valleys indicate the font size used in the written text and the distances between the lines.

Figure 5 provides an example of vertical and horizontal projections of a text region. These projections are normalized by the height and width of the image, respectively. The RLE technique is applied to the projection vectors where the mean and standard deviation (STD) of the resulting RLE coefficients are computed. If the paragraph is written in a consistent font and the spacing between its lines is fixed, this will generate a relatively low STD value in comparison with the average line width indicating a text region. Therefore, if the average line width is higher than the variation (STD) at least in one direction, the image region is identified as a text region. An example is shown in Fig. 5 where the pattern (peaks and valleys) is formed in the horizontal projection [see Fig. 5(c)].

2.3 *Photo-Detection Module*

Similar to the text-detection module, the preprocessed L* channel is used as an input for the photodetection module. It is initially segmented into two classes, background and photo, by utilizing projection basis vectors. After an initial segmentation is achieved, an MRF-MAP optimization with iterative conditional modes (ICM) is employed to segment neighboring/local spatial information and merge the outcomes in a more accurate photo map. As a last step, missing blocks (false positives) and misclassified blocks (false negatives), which are fully surrounded by detected photo block (s), are included in the final photo map. A block diagram of the photo detection module is shown in Fig. 6, and the corresponding submodules are explained in the following subsections.

2.3.1 *Block-wise segmentation based on basis vectors projection*

Assuming the document is skew-corrected, letters or words are printed on a document from left to right or top to bottom with some predetermined spaces. The combination of the spaces and words provide specific patterns. In this step, these specific patterns are utilized to eliminate background and text regions and obtain an initial photo map for a given input scanned document. The initial photo map is generated using projection basis vectors by processing the given document in a block-by-block fashion.

Block-wise segmentation based on projection basis vectors was first proposed by Won,[35] where a scanned document

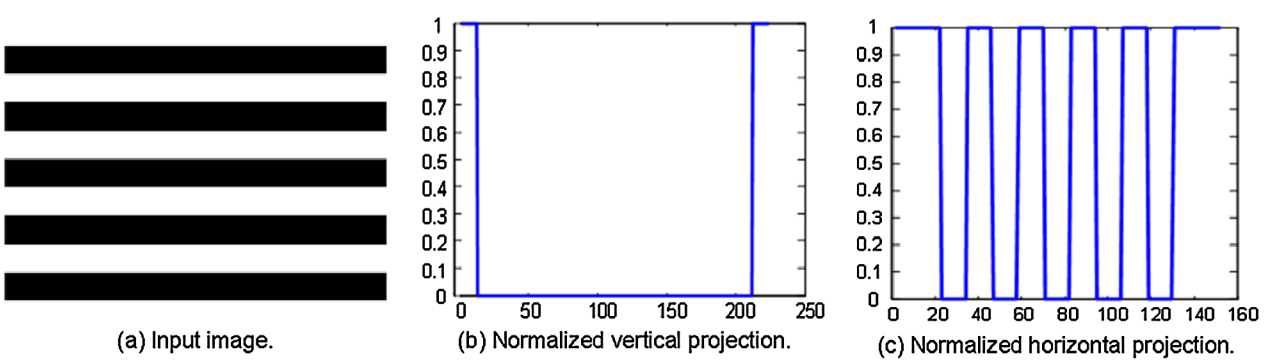


Fig. 5 Example of vertical and horizontal projections of a text region.

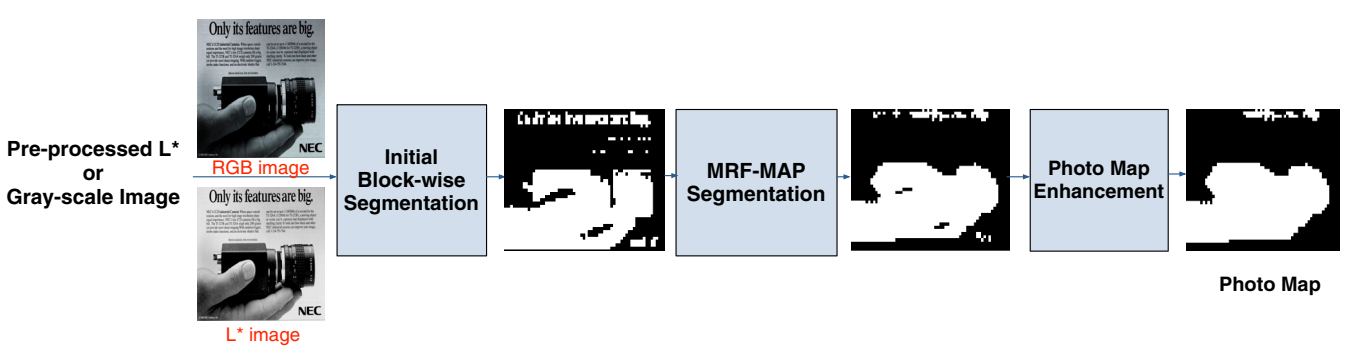


Fig. 6 Block diagram of photo-detection module.

is directly employed as an input image. In other words, no preprocessing is applied to the input RGB image. Additionally, the methodology introduced in Ref. 35 determines the optimum block-size by utilizing alternating blanks between words and text lines. The constraint for an optimum block-size is to include sufficient number of text lines in the block. On the contrary, in our study, the block-size $B \times B$, is fixed to 32×32 pixels since the proposed technique is employed to classify only photo regions in the scanned document. However, if any down-sampling is applied to the input image in the preprocessing module, an updated block-size is obtained by multiplying the $B \times B$ pixels window size by the corresponding scaling factor. In addition, the decision criterions for background and photo regions are modified to yield a more robust initial photo map for different types of complex color background scanned documents [see Eq. (8)].

Initially, the block-wise segmentation is achieved by utilizing projection basis vectors. These different types of basis vectors represent either text, background, or photo regions (although the technique is not specifically applied to detect text regions in the document). The set of the pixel locations, defined as $\Omega = \{(i,j) : 0 \le i \le N_1, 0 \le j \le N_2\}$, is divided into $B \times B$ nonoverlapping blocks. Then the image becomes a set of blocks whose indices are denoted as $\Omega_B = \{(k,l) : 0 \le l \le n_1, 0 \le l \le n_2)\}$ where n_1 and n_2 are $\lfloor N_1/B \rfloor$ and $\lfloor N_2/B \rfloor$, respectively. Furthermore, a scanned document data field (a set of image pixels) is defined as $y = \{y_t(i,j) : t \in \Omega_B\}$ and $y_t = \{y_t(i,j) : 0 \le i,j \le B-1\}$. First, for each block, the gray-levels in preprocessed L* image are horizontally projected in order to constitute a row-vector $H_t = [h_t[0], h_t[1] \cdots, h_t[B-1]]^T$, where H_t represents the projection values for a horizontal line in the selected block. $h_t[k]$ takes either a value of $+1$ or -1 depending on whether the k'th line corresponds to a nonbackground or background as shown in Eq. (3):

$$h_t[k] = \begin{cases} +1 & \text{if } \left[\sum_{j=0}^{B-1} I\{y_t(i,j)\}\right] > B * T_2 \\ -1 & \text{if otherwise} \end{cases},$$

$$\text{where } X\{y_t(i,j)\} = \begin{cases} 1, & \text{if } y_t(i,j) > T_1 \\ 0, & \text{if otherwise} \end{cases}, \tag{3}$$

where $k = i\,(\mathrm{MOD}\,B)$ corresponds with the number of lines in the selected block and $I\{y_t\}$ represents the pre-processed L* image pixels that are binarized according to the threshold value T_1. Then, if the corresponding line (k) in Eq. (3) represents a nonbackground line, then most of its pixels will have lower values than $B * T_2$ and $h[k]$ takes the value of -1. Otherwise, they will be classified as background and $h[k]$ will be assigned to 1. Subsequently, the number of 1's in each horizontal line of the block are counted to classify the block as part of the nonbackground or background zone.

The basis vectors, shown in Fig. 7, are utilized to select the appropriate class for the block of interest. Note that the vectors in Fig. 7 are orthonormal in order to span the space and satisfy the normalization conditions as formulated in Eq. (4).

$$\langle \Phi_i, \Phi_j \rangle = \sum_{k=0}^{7} \phi_{ik}\phi_{jk} = 0 \quad \text{and}$$

$$\langle \Phi_i, \Phi_i \rangle = \sum_{k=0}^{7} \phi_{ik}\phi_{ik} = 1, \tag{4}$$

where $0 \le i,j \le 7$ for $\forall\ i,j$ and $\langle . \rangle$ represents the inner product of any two basis vectors. Φ_0 and Φ_1 illustrate background and photo region patterns, respectively, while the rest represents different type of text patterns. To satisfy the dimensional compatibility in vector multiplication given in Eq. (7), the horizontal lines in the block are divided into eight groups by utilizing Eq. (5):

$$\hat{h}_t[k] = \sum_{n=0}^{[(k+1)B/8]-1} h_t[n], \tag{5}$$

where $k = 0, 1, \cdots, 7$. Hence, the rearranged vector becomes $H_t = [h_t[0], h_t[1], \cdots, h_t[7]]$. Moreover, H_t can be expressed using the basis vectors with some weighting coefficients given by Eq. (6):

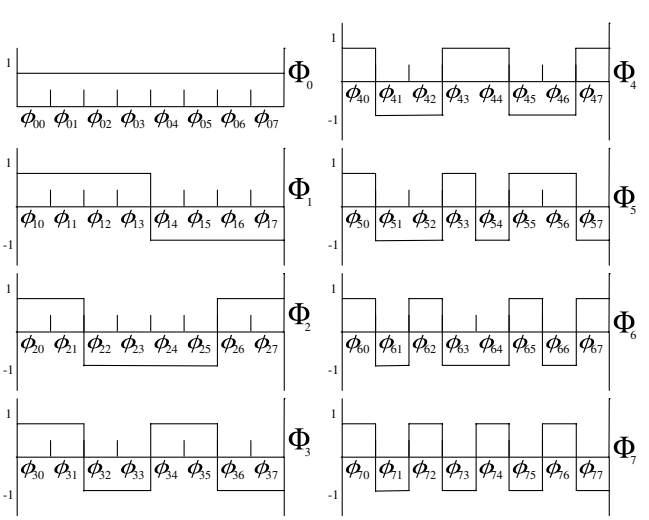


Fig. 7 Basis vectors for the determination of the best fit for the region in the block of interest.[35]

$$\hat{H}_t = \frac{1}{8}[a_t[0]\Phi_0 + a_t[1]\Phi_1 + \cdots + a_t[7]\Phi_7] = \frac{1}{8}\sum_{k=0}^{7} a_t[k]\Phi_k, \tag{6}$$

where $t = 0, \cdots, 7$ and $a_t[k]$ represents the weighting coefficient that corresponds to the basis vector, Φ_k. Similar to the discrete-time Fourier series, $a_t[k]$ can be obtained by the inner product of H_t and Φ_t as shown in Eq. (7):

$$a_t[k] = \langle \hat{H}_t, \Phi_k \rangle = \hat{h}_t[0]\phi_{k0} + \hat{h}_t[1]\phi_{k1} + \cdots + \hat{h}_t[7]\phi_{k7}, \tag{7}$$

where $k = 0, \cdots, 7$. To finalize the initial block-wise segmentation, class labels $\hat{x}_t$, are assigned as 2 for background block, 1 for photo block, and 0 for text block according to Eq. (8):

$$\hat{x}_t = \begin{cases} 2 & \text{if } |a[0]| \geq |a[l]| \quad \text{and} \quad |a[0]| > \sum_{i=1}^{7} |a[i]| \quad \text{and} \quad a[0] > 0 \\ 1 & \text{if } |a[0]| \geq |a[l]| \quad \text{and} \quad |a[0]| > \sum_{i=1}^{7} |a[i]| \quad \text{and} \quad a[0] < 0 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where $l = 1, \cdots, 7$. If $|a[0]| \geq |a[l]|$ and $|a[0]| > \sum_{i=1}^{7} |a[i]|$, then the first coefficient $a[0]$, becomes the most dominant coefficient. In addition to this condition, if $a[0] > 0$, the pixels values in the selected block have the tendency to have monotone levels, corresponding to a background block. On the other hand, if $|a[0]| \geq |a[l]|$ and $|a[0]| > \sum_{i=1}^{7} |a[i]|$, then the first coefficient again becomes the most dominant coefficient. However, in this case, if $a[0] < 0$, then the pixels values in the selected block tend to have both monotone and nonwhite levels, which represent a photo block. Otherwise, the block corresponds to a text block which consists of a set of horizontal gaps between the words which are mimiced by the basis vectors.

2.3.2 *Segmentation based on MRF-MAP optimization*

Contextual information is not considered while developing the initial block-wise segmentation stage described in Sec. 2.3.1. Therefore, the block-wise segmentation result $\hat{x}_t$ has possible false detections. To eliminate these misclassifications, a block-based MRF-MAP segmentation technique in an ICM framework is employed. It is a deterministic algorithm that maximizes local conditional probabilities, denoted by $p(y|x)$, iteratively based on two basic assumptions. The first is that neighboring pixels tend to have the same values since images consist of regions that are highly correlated except around edges. This provides an opportunity to change a pixel label that is corrupted by noise, based on its neighborhood information. The second assumption is that each pixel is corrupted by an independent identically distributed Gaussian noise. In this regard, the initial segmentation result $\hat{x}_t$, obtained in Sec. 2.3.1, is considered as initial photo map of the photo class label $X = x$ where $x = \{x_t = \hat{x}_t : t \in \Omega_B\}$. The scanned document data is assumed to be a random field $Y = \{Y_t : t \in \Omega_B\}$, and it is equal to the gray-levels in preprocessed L* image in each block of Ω.

The goal of a MRF-MAP framework is to maximize the a posteriori probability density function (pdf) of the segmentation field, given by Eq. (9).

$$p(x|y) \propto \arg\max_x p(x)p(y|x), \tag{9}$$

where $p(x)$ is *a priori* pdf of the region assumed to be Gibbs distribution as defined in Eq. (10):

$$p(x) = \frac{1}{Z}\exp\left[-\sum_{c\in C} V_c(x)\right], \quad \text{where } V_c(i,j;k,l) = \begin{cases} -\beta & \text{if } x(i,j) = x(k,l) \\ +\beta & \text{otherwise} \end{cases}, \tag{10}$$

where Z is the normalization constant, $V_c(x)$ is the clique potential, and C is the set of all cliques in $B \times B$ block. "Second order neighborhood clique system" (3×3 pixels sized window), first introduced by Derin et al.[39] is utilized to obtain the clique potentials $V_c(x)$, while computing a priori probability $p(x)$. The two-block clique potentials $V_c(i,j;k,l), (i,j), (k,l) \in c$, are also given in Eq. (10). The clique potential constant β is chosen empirically to be 1.6.

The second term in Eq. (9) $p(y|x)$, is the conditional pdf of the pre-processed L* image data given in Eq. (11). To find the class conditional pdf, each $B \times B$ nonoverlapping block is assumed to be independent and modeled by Gaussian pdf given the class label field $X = x$ since the mean and variance are sufficient statistical features to characterize the blocks. As a result, $p(y|x)$ can be written as:

$$\begin{aligned} p(Y = y|X = x) &= \prod_{(i,j)\in\Omega_B} p[Y(i,j) = y(i,j)|X(i,j) = x(i,j)] \\ &= \prod_{(i,j)\in\Omega_B} \frac{1}{\sqrt{2\pi\sigma^2_{x(i,j)}}}\exp\left\{\frac{-[y(i,j) - \mu_{x(i,j)}(i,j)]^2}{2\sigma^2_{x(i,j)}}\right\}, \end{aligned} \tag{11}$$

where Ω_B represents the image and $\mu_{x(i,j)}(i,j)$ and $\sigma^2_{x(i,j)}$ are the mean and variance function for each distinct class as a function of (i,j), respectively. After defining each variable by using Eqs. (10) and (11), the expression in Eq. (9) can be simplified while omitting the normalization constants $1/Z$ and $\sqrt{2\pi\sigma^2_{x(i,j)}}$ as follows;

$$p(x|y) \propto \arg\max_{x} \times \left\{ \exp\left[-\sum_{(i,j)\in\Omega_B} \frac{[y(i,j)-\mu_{x(i,j)}(i,j)]^2}{2\sigma^2_{x(i,j)}} - \sum_{c\in C} V_c(x) \right] \right\}, \quad (12)$$

where the first summation term constraints the region intensity to match the available data and the second imposes spatial continuity. The expression in Eq. (12) computes the energy over the entire image, which is computationally very expensive. Hence ICM is utilized to strike a balance between accuracy and computational complexity. To this effect, instead of considering the entire image, $p(x|y)$, which corresponds with an energy in a given block/window, is computed for a window size of 32×32 pixels [$i = j = 32$ in Eq. (12)] employing the same clique system while finding *a priori* probability $p(x)$.

This iterated approach is executed until the convergence condition, denoted by CC and given below, is satisfied. Therefore, the algorithm ceases once the number of updated class labels $x'(i,j)$, is less than 10% of the overall image. The aforementioned module can be summarized in the following steps:

1. For given current class labels x, calculate (μ_0, σ_0^2), (μ_1, σ_1^2), and (μ_2, σ_2^2), which represent the mean and variance of text, photo, and background zones, respectively.
2. Compute $p(x|y)$ for each block in the image given in Eq. (12) and update the current class labels of the blocks x', by selecting the class (0, 1, or 2), which maximizes the energy for a given block.
3. If $CC < T = 0.1$, stop. Otherwise, go to step 1.

$$CC = \frac{1}{B^2} \sum_{i=0}^{B-1} \sum_{j=0}^{B-1} \text{sign}[x'(i,j) - x(i,j)],$$

where

$$\text{sign}[x'(i,j) - x(i,j)] = \begin{cases} 1 & \text{if } x'(i,j) \neq x(i,j) \\ 0 & \text{otherwise} \end{cases}, \quad (13)$$

Note that the result of the summation in Eq. (13) will change if only if $x'(i,j)$ and $x(i,j)$ have different class labels. The algorithm overly converges in 2 or 3 iterations.

2.3.3 *Photo map enhancement process*

Once the MRF-MAP based image segmentation is completed, a photo map enhancement step is performed to eliminate false negatives [shown as black pixels which are surrounded by classified blocks (white pixels) in Fig. 8(a)], and to generate a final photo map as illustrated in the synthesized image in Fig. 8(b).

A second-order neighboring connectivity model is used along with morphological dilations in an iterative fashion until the contour of the initial subband fits within a primary detected region (small and big rectangular areas in Fig. 8). The process stops when further dilations may result in potential changes in the shape (contour) of the main detected image. On the other hand, if a false negative zone has a connection with the main detected nonimage region, that zone is not joined to the main detected subband as can be seen in Fig. 8(b). Finally, as done in the text-detection module, regions with size less than 0.03% of the input scanned document size are removed from the binary photo-candidate map.

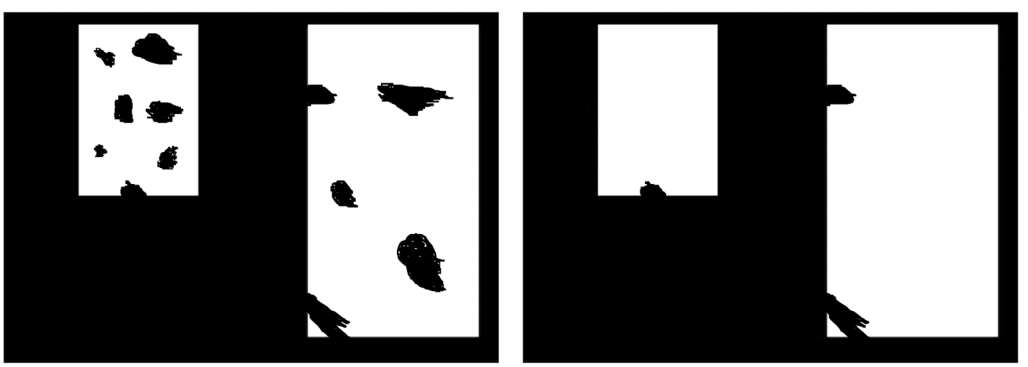
2.4 *Strong Edge/Line Detection Module*

Lines are detected in the proposed scheme using Hough transform. The algorithm starts by employing Canny edge detection methodology to generate an edge map of the input preprocessed L* channel followed by the Hough transform using the parametric representation of a line, given in Eq. (14) below:

$$\rho(\theta) = x\cos(\theta) + y\sin(\theta). \quad (14)$$

The Hough transform generates a parameter space matrix whose rows and columns correspond with ρ and θ, respectively. Peak values in this plane represent potential lines in the input image. Several parameters that are essential for the success of the line detection algorithm are set empirically based on our extensive training data-set. These are given as follows:

1. A threshold value equal to 20% of the maximum peak is used to identify potential lines.
2. The maximum number of peaks to identify in the parameter space matrix is set to 30.



(a) Before photo map enhancement process. (b) After photo map enhancement process.

Fig. 8 Photo map enhancement.

3. A scalar value that specifies whether merged lines should be kept or discarded is set at 300 pixels such that lines shorter than this value are discarded.
4. When the distance between two line segments associated with the same Hough transform bin is less than 15 pixels, the Hough methodology merges the line segments into one single line-segment.

The strong edge/line detection module takes an edge map, generated by the Canny edge detection algorithm, as input. Edge pixels are linked together into lists of sequential edge points; one list for each edge contour. A contour or edge-list starts/stops at an ending or a junction with another contour/edge-list. A thresholding technique is used to eliminate short edges where contours less than 300 pixels long are discarded.

2.5 *Map Combination*

The map combination module is employed to classify the pixels/blocks that are identified in both photo and text maps. The goal is to classify these intersected pixels/blocks as either a photo or a text region as shown in Fig. 9. Areas included in both text and photo maps are utilized to generate the intersection map, while the rest of the text and photo regions are used as training maps. Three low-level vision features are estimated from the training maps and used to classify the intersection regions. A block diagram of the map combination module is shown in Fig. 10.

Three local image features are employed in the map combination module, namely: (a) standard deviation (STD) in the horizontal direction, (b) standard deviation in the vertical direction, and (c) the region entropy. Standard deviation in the horizontal and vertical directions are computed by dividing the training maps into blocks of the same window size as discussed in block-wise segmentation, (32×32). Each window gives a STD value in the horizontal direction and another in the vertical direction. Thus entire block (intersection map) forms a vector for each training map. As expected, horizontally oriented text regions tend to possess higher standard deviation values than photo regions (background and text pixels form a relatively higher contrast compared to the photo regions). STD in vertical direction is also considered to accommodate vertically oriented text. Although photo regions may display a noted contrast with respect to the background, it is generally not as high as text regions.

Entropy is a statistical measure of randomness that can be used to characterize the texture of the input image. It is defined as follows:

$$\mathrm{Ent}_R = -\sum_{i=0}^{B-1}\sum_{j=0}^{B-1} q \log_2(q), \tag{15}$$

where q is a vector, which contains the probabilities of each gray level that appears in the input image, and R is the target region that is either text or photo. The q vector can be easily obtained by utilizing the image histogram. According to our experiments, text regions tend to possess a higher entropy value (higher degree of randomness) than photo regions.

Once the features are computed, K-Means clustering is employed to minimize the Euclidean distance, resulting in a final segmentation map. This module is skipped if there is no intersection map, or the data in the training maps is not sufficient to compute the features. Three different features are employed to separate between the text and photo regions in a clustering framework. The coordinates of the centroids are the average of STD in both directions and entropy. The algorithm is initialized with the computed centroids. Each block in the intersection map is classified

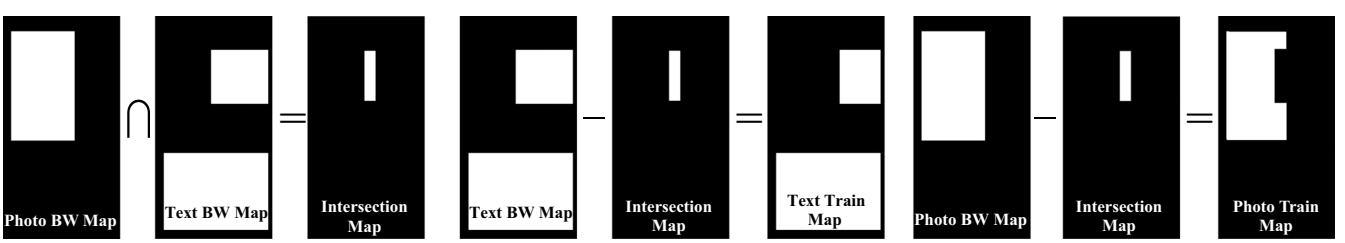


Fig. 9 Define training and intersection maps.

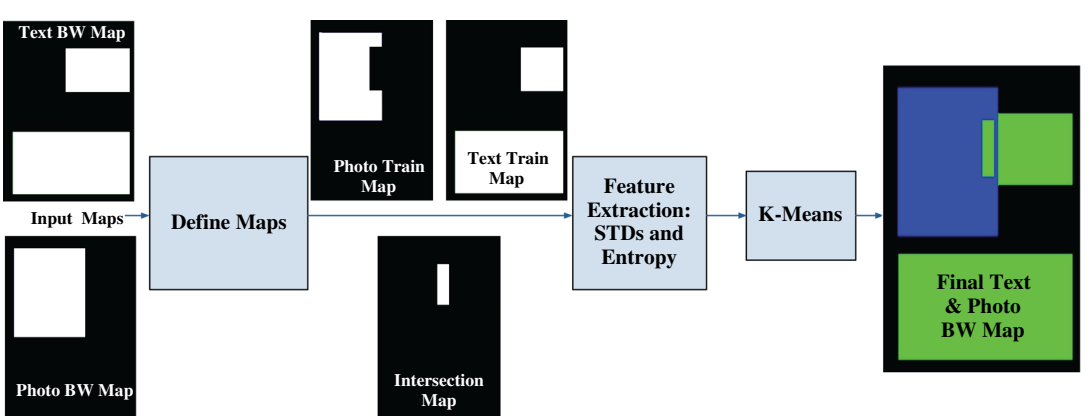


Fig. 10 Block diagram of the map combination module.

as either a text or photo region. However, some of the class members are misclassified since the classes are not linearly separable. To reduce the computation time, centroids are not updated as normally done in *K*-Means, since the amount of training data used for extracting features is much larger than the test data. The iterative procedure is terminated when blocks in the intersection map are completely classified. The regions in blue and green color in Fig. 10 represent the final detected photo and text regions, respectively. Strong edge/line map generated by the corresponding module is not considered in this module. Instead, it is directly combined with the resultant photo and text map after the intersection map is completely processed since the strong edge/line map does not have a considerable intersection zone(s) with the text and/or photo map.

3 Results and Discussions

The proposed scheme was tested on the Oulu University MediaTeam document database,[40] which contains a variety of simple to complex color and gray-scale scanned documents such as *Articles, Advertisements, Newsletters, Business-cards and Dictionary.* In Ref. 40, music and line drawings are considered line art rather than photo documents and thus are excluded from our study. The resulting page layout classification map outlines text and photo regions by utilizing corresponding rectangular boxes. On the other hand, the detected lines and edges are shown as an edge map. It is worth noting that our technique was designed with the capability to produce both pixel-wise and box-wise output maps. However, the Oulu database provides only box-wise ground-truth data. Hence our page classification results were compared against the Oulu database text and photo provided ground-truth maps. The accuracy rates of the aforementioned comparisons are presented in Tables 1 and 2.

In our algorithm, the preprocessing module is utilized to enhance the performance of the text and photo classification especially in the presence of complex color backgrounds. Its significance is clearly illustrated in Fig. 11, where the proposed method is applied with and without the preprocessing module on two different (correspondence and advertisement) types of scanned documents. Although the input image in Fig. 11(a) has a photo region in its background, this region should not be included as a photo in the classification map. However, our technique incorrectly detects the aforementioned region and classifies it as a photo when it is employed without the preprocessing step. On the other hand, the classification map is clearly more accurate when the preprocessing module is utilized. Similar result is achieved in Fig. 11(b) using a typical magazine page that contains a complex background. Applying the preprocessing module eliminates the background from the classification map and yields to a satisfactory outcome.

Several documents from Ref. 40 were also selected to benchmark the performance of the proposed technique quantitatively utilizing a confusion matrix (CM) approach. However, since the Oulu database[40] does not provide ground-truth for strong-edges and lines, these could not be presented in CM and were shown qualitatively. Hence the results for line and strong edge classification are instead illustrated in Figs. 12 and 13 below as pixel- and box-wise maps.

Figure 12 illustrates the generated page classification maps for two documents that contain text, lines, and strong-edges/lines. The original color documents are shown in Fig. 12(a). The color space conversion (RGB to CIE-L*a*b*), applied in the proposed algorithm, eliminates artifacts in background regions as shown in Fig. 12(b).

Table 1 Confusion matrix for 16 different types of scanned document.

	Proposed algorithm		
Ground-truth	**Text**	**Photo**	**Background**
Text	87.12	1.69	11.19
Photo	2.69	91.22	6.09
Background	3.89	0.67	94.44

Table 2 Performance comparison between Duong et al.[23] and the proposed algorithm.

Document class	# of samples	Av. Perf. [23]	Proposed algorithm Av. Perf.
Address-list	6	0.75	0.81
Advertisement	24	0.95	0.91
Article	233	0.75	0.88
Business cards	11	0.96	0.91
Check	3	0.93	0.81
Color-seg-images	10	N/A	N/A
Correspondence	24	0.82	0.91
Dictionary	12	0.97	0.95
Form	23	0.86	0.82
Manual	35	0.88	0.87
Newsletter	42	0.86	0.89
Outline	19	0.84	0.80
Phone book	7	0.88	0.93
Program listing	12	0.92	0.78
Street map	3	1.00	0.87
Terrain map	5	0.93	0.90
Math	17	0.67	0.78
Music	9	0.84	N/A
Line drawing	7	0.95	N/A

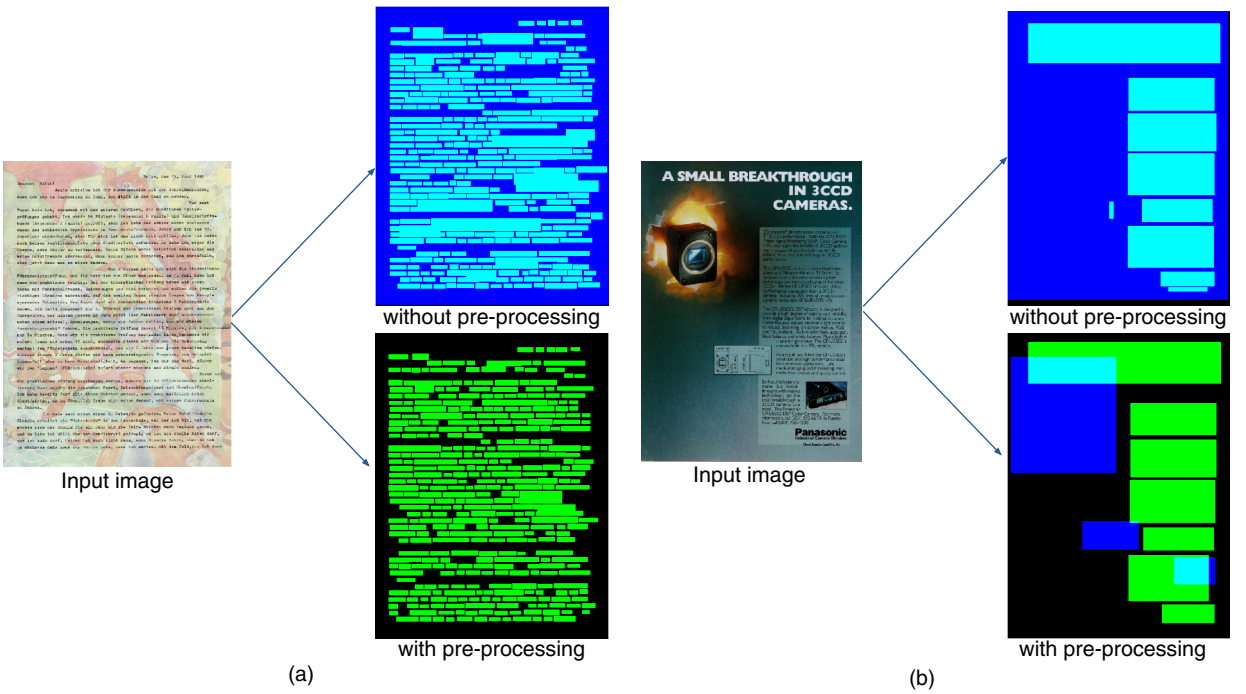


Fig. 11 Input scanned documents: (a) correspondence and (b) advertisement/magazine.

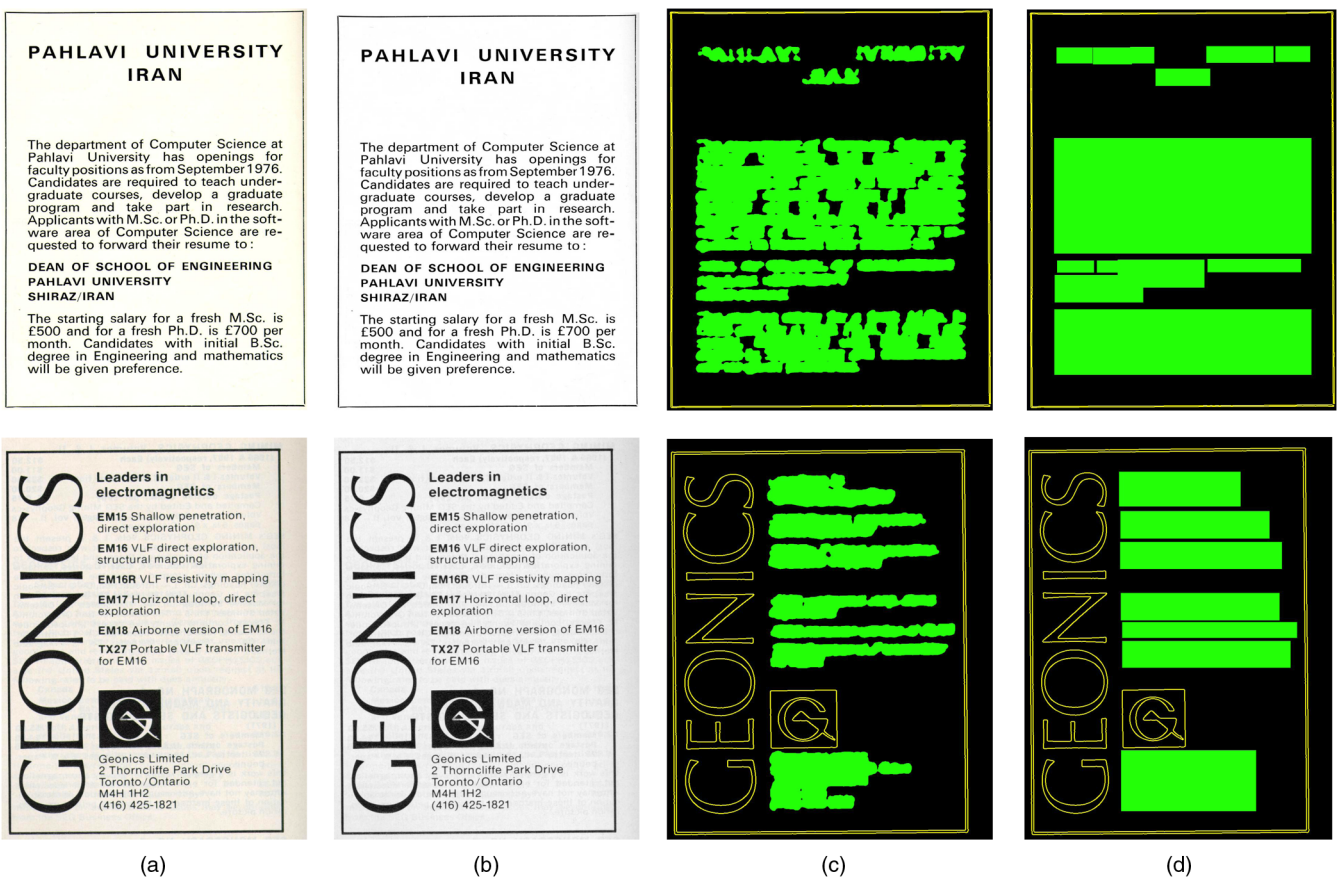


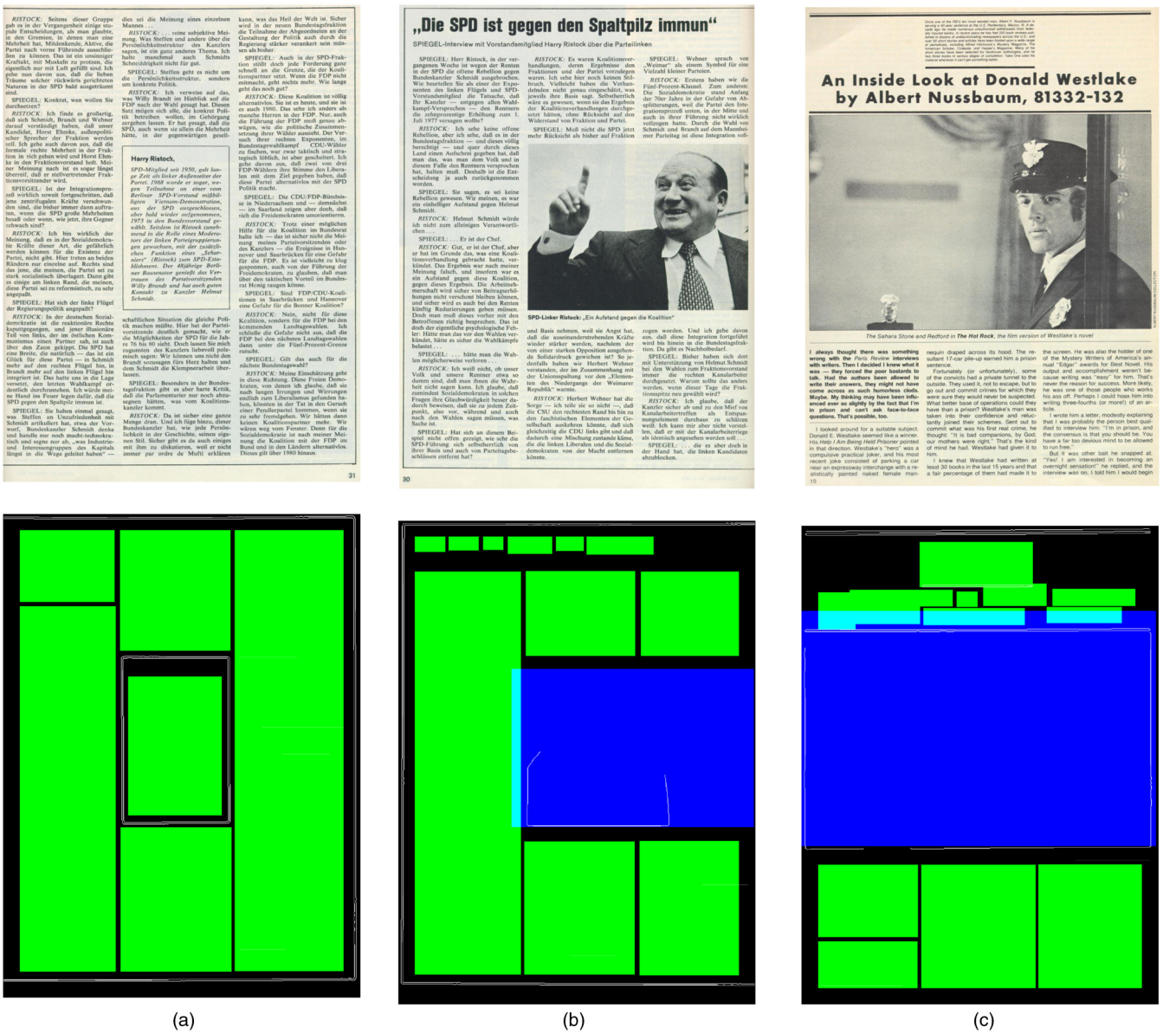
Fig. 12 Results for line detection: (a) original image, (b) preprocessed L* channel, (c) pixel- and (d) box-wise final classification map.

Furthermore, the enhanced document enables better detection accuracy as demonstrated in Fig. 12(c) and 12(d) where strong edge/line and text regions are colored in yellow and green, respectively. Notice that the documents shown in Fig. 12(a) and 12(b) have frames (box-lines) that outline the pages. These aforementioned frames are accurately detected in both images as can be seen in the figures in addition to the written text with larger font size, which is detected as strong-edges as well (second row of Fig. 12). Moreover, if the spaces between words are significantly noticeable, the pixel-wise classification can detect these spaces as shown in Fig. 12(c). The pictorial structure shown in the document is also accurately detected.

In Fig. 13, three different sample documents are utilized to demonstrate the effectiveness of the strong edge/line detection method on scanned documents that have both photo and text zones. Notice that the frames of the first document [see Fig. 13(a)] are well-extracted with the text regions.

Furthermore, the performance of our algorithm is clearly demonstrated on a series of RGB and gray-scale type documents as shown in Figs. 14–16. In each figure, the photo and text regions are represented in blue and green, respectively, in each of the ground-truth and classification maps. Overlapping photo and text regions are shown in cyan.

Figure 14 provides a set of challenging documents of various types; namely address list [see Fig. 14(a)], advertisement [see Fig. 14(b)], article [see Fig. 14(c)], business card [see Fig. 14(d)], and check [see Fig. 14(e)], along with their corresponding classification maps and ground-truth data. In each of these cases, our results indicate that the proposed algorithm performs extremely well when compared with the human-generated ground-truth maps with minor exceptions: 1. the tiny region (false positive) in the photo region detected as photo and text in address list, 2. the misclassified figure captions at the top and bottom of the page in article, and 3. the three horizontal-lines region (false positive), which lies at the left of business card owner's



(a) (b) (c)

Fig. 13 Results for line detection: document (a) w/o photo, (b) with photo and (c) with strong edge/line, text and photo.

in business card. It should be noted that our technique furnishes a more accurate classification of the text-regions in business card and check in comparison to the ground-truth data, which tends to group these into a single box.

Figure 15 shows a set of documents with nearly identical classification results for both color and gray-scale scanned originals. Similarly to what was described earlier, our system provides highly accurate classification results that correlate well with the human-generated ground-truth for both photo and text regions with a few minor discrepancies: 1. a small part of the photo region is misclassified as text in correspondence since the text region at the left of the photo is segmented in the same box as the top of the photo as shown in Fig. 15(b); 2. parts of the second table are miss detected in the form document as illustrated in Fig. 15(d); and 3. our scheme shows detailed text regions in the manual document while the ground-truth map masks the entire lower region in Fig. 15(e) as a single text area.

The last set of scanned documents, which consists of newsletter, outline, phone book, street map, and terrain map, is presented in Fig. 16. The results, once again, clearly emphasize the accurate performance of the proposed algorithm with the following minor exceptions: 1. the photo region at the top of the page (circle with stars) in newsletter is misclassified since the main body of the region includes background [see Fig. 16(a)], and 2. the small text regions corresponding to the country or street names in Fig. 16(d) and 16(e). Notice that the body of the text zone in the outline scanned document in Fig. 16(b) is accurately detected, and the photo module manages to clearly differentiate between a real photo region and the gray part of the background at the right side of the document. Moreover, the text regions in Fig. 16(c) are well-segmented, and the photo detection module effectively differentiates the complex background and any photo region for both color and gray-scale scanned documents. Similarly, in street map and terrain map, the maps are correctly classified as photos.

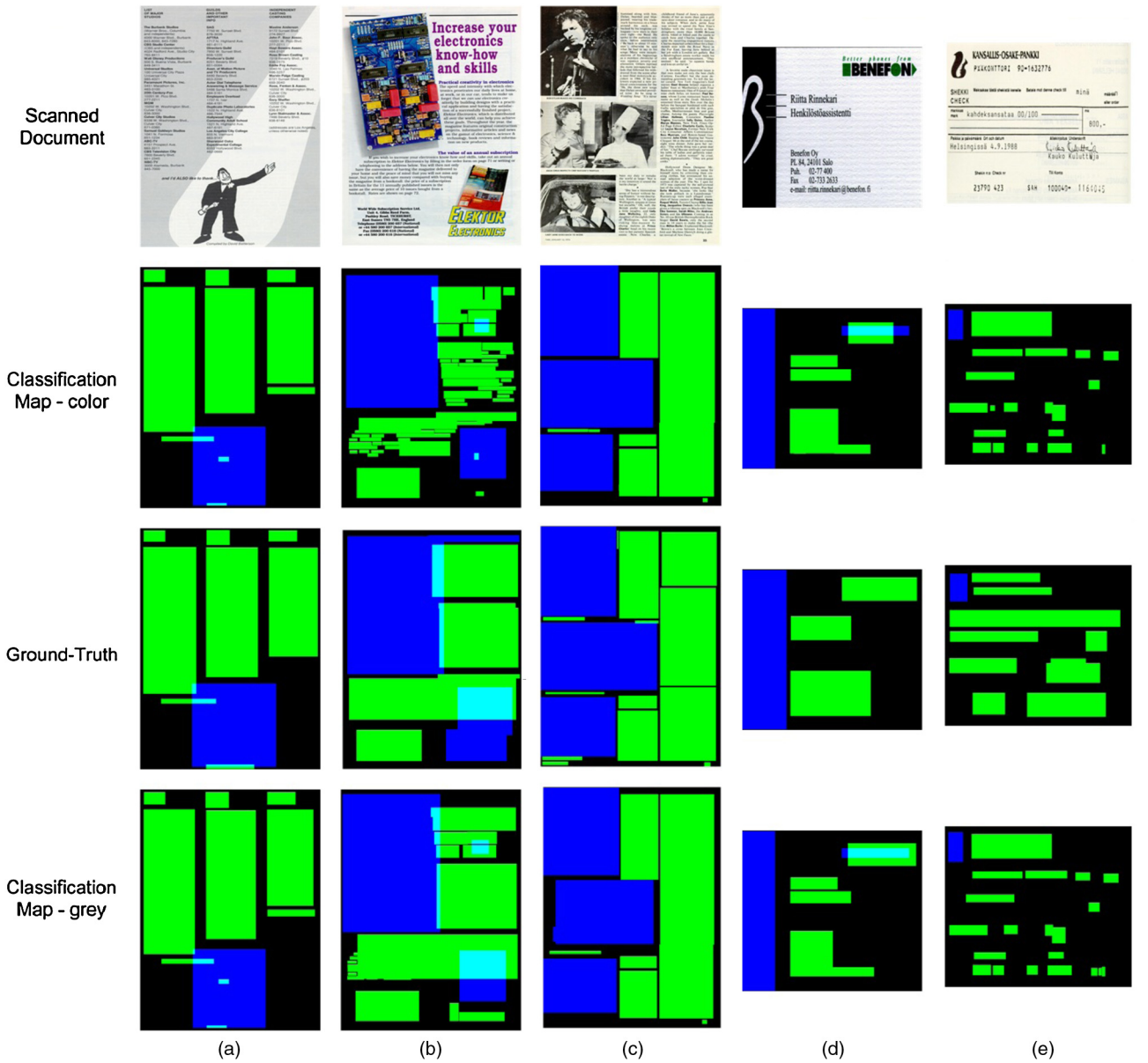


Fig. 14 Final classifications map for (a) address list, (b) advertisement, (c) article, (d) business card, and (e) check scanned document.

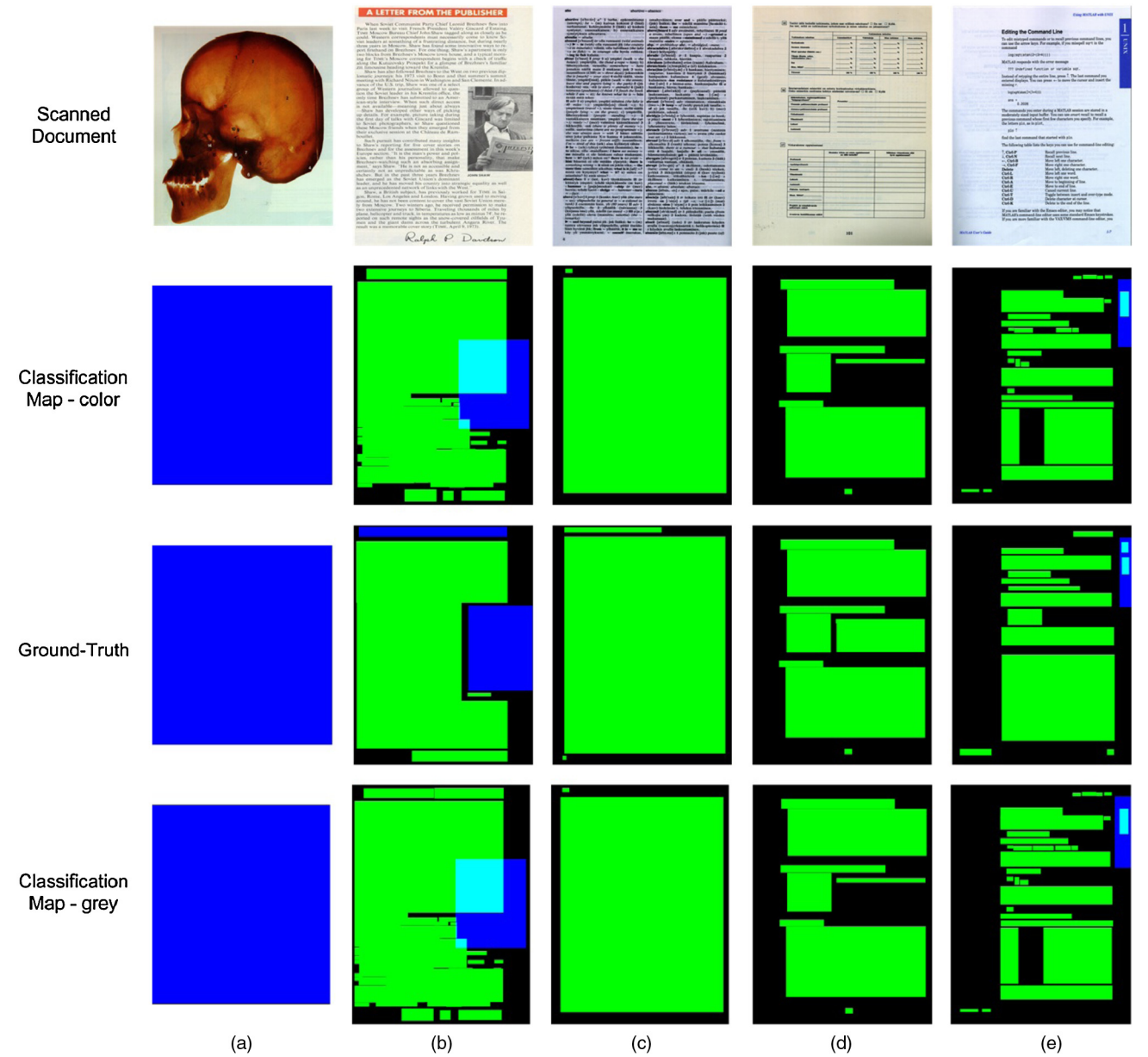


Fig. 15 Final classification map for (a) color Segmentation, (b) correspondence, (c) dictionary, (d) form, and (e) manual scanned document.

A quantitative analysis of the proposed system's performance is presented in Table 1 and Fig. 17. Table 1 illustrates the average classification accuracy rates in comparison with the database ground-truth data[40] using a confusion matrix. The parametric values used in the photo detection module, namely $\beta = 1.6$, $T_1 = 200$ and $T_2 = 0.7$, are fixed [see Sec. 2.3.1]. B is set to 32 and 16 for document images with sizes comparable to 3000×2000 pixels and 2000×1000 pixels, respectively. In summary, Table 1 shows that 87% classification accuracy is found in text regions with 11% misclassified as background. Conversely, a classification rate of 91% and 94% is achieved for photo and background detection, respectively.

Individual categories of documents are utilized to test the algorithm as shown in Fig. 17. Accuracy rates for advertisement, article, correspondence, and newsletter document types are calculated individually. However, the document type MOD is made of 35 manual, 19 outline, and 12 dictionary documents. MOD is formed to obtain a sufficient data-set and evaluate the performance of the proposed methodology. The rest of the database, totaling 97 documents, is utilized to compose the document type OTHER.

3.1 *Performance Evaluation*

In this section, the performance of our technique is compared with two state-of-the-art document classification algorithms using Oulu database.[40] Both methods are applied to scanned documents of size $\sim 3000 \times 2000$ pixels. The first one was introduced by Duong et al.[23] where the entire Oulu database is used to test for textual region detector. The second was developed by Won,[35] where 233 article documents were used to evaluate the image extraction system. Both techniques in Refs. 23 and 35 used pixel accuracy rate as a performance metric.

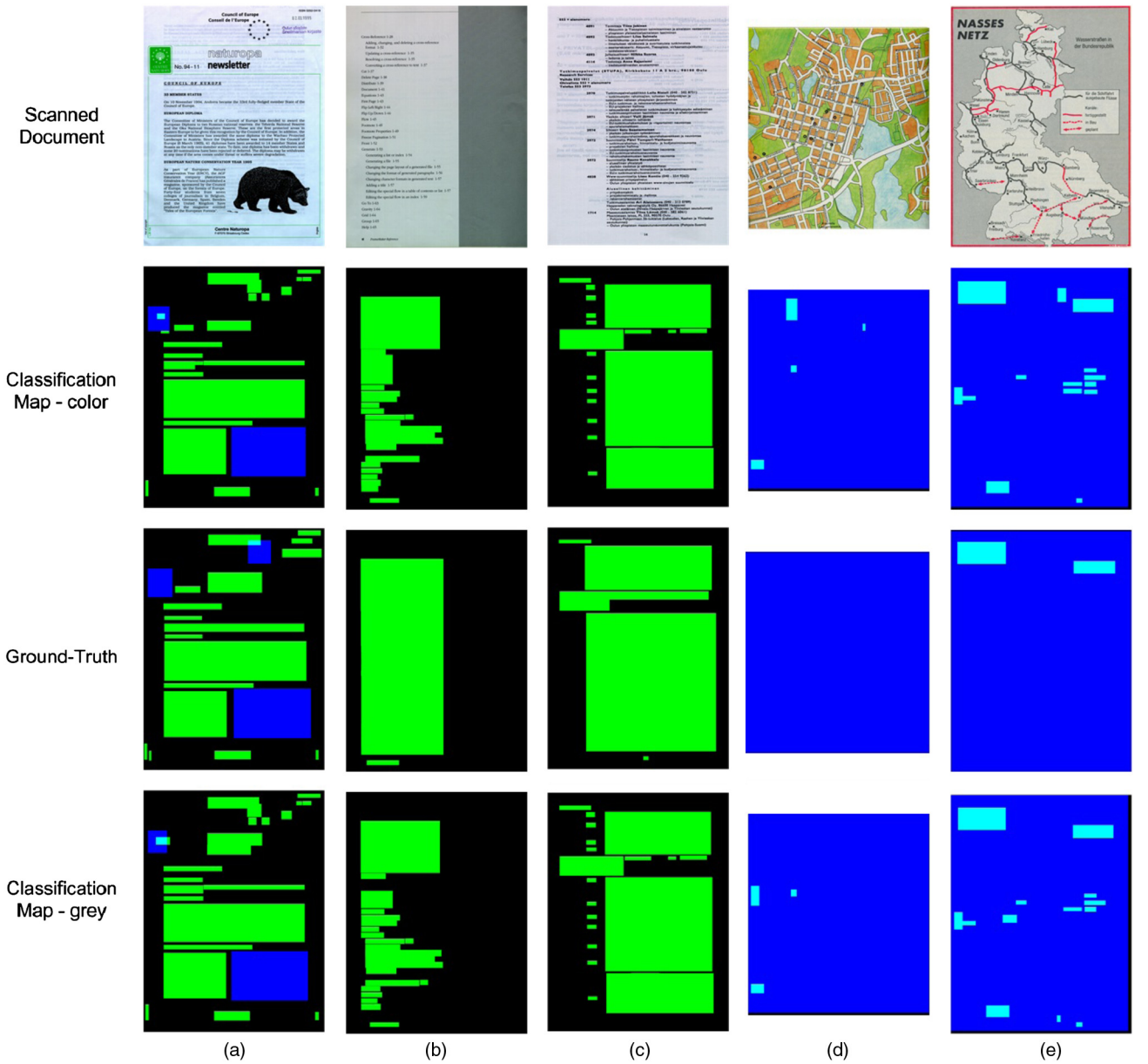


Fig. 16 Final classification map for (a) newsletter, (b) outline, (c) phone book, (d) street map, and (e) terrain map scanned document.

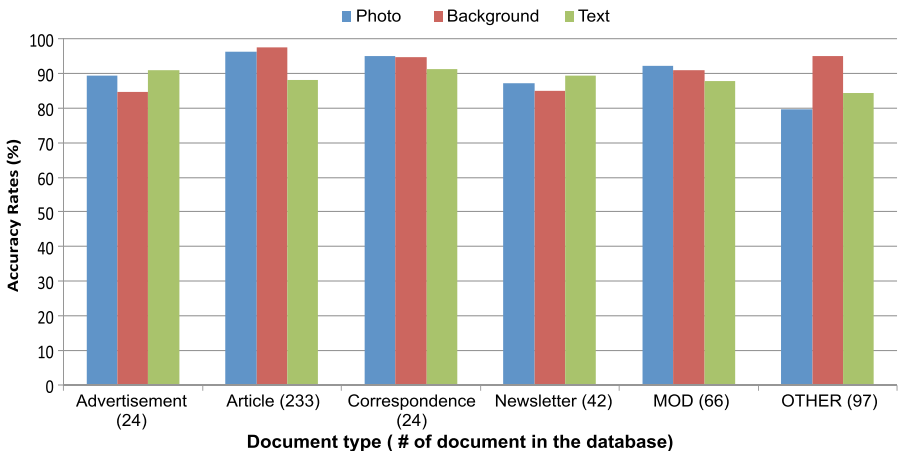
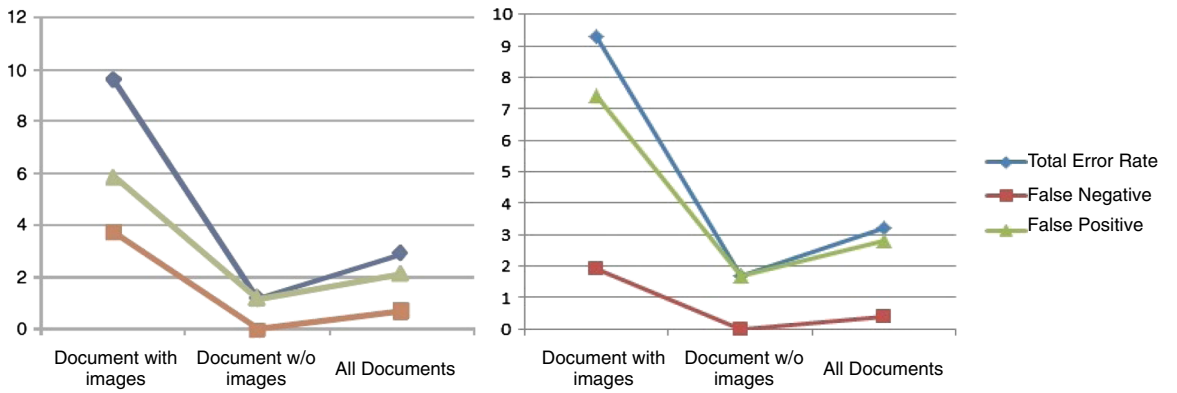


Fig. 17 Accuracy rates for photo, text, and background classes in different types of scanned documents.



(a) Won [35] (b) Proposed algorithm

Fig. 18 Error rates(%).

3.1.1 *Comparison to work done by Duong et al.*

Duong et al. propose a two-step document analysis system that detects regions using cumulative gradient considerations and classifies them as text and non-text zones utilizing geometric and texture features.[23] The entire MediaTeam document database,[40] which includes ~500 document images, is tested to validate the performance of their system. Text regions are extracted approximately as rectangular areas defined by their bounding boxes. In Ref. 23, the accuracy rates are computed using Eq. (16):

$$\text{Accuracy}(\%) = \frac{|t|}{|T|}, \tag{16}$$

where T is the set of the rectangular text regions defined by the database, and t is the set of text regions segmented successfully by the proposed system. The accuracy rate comparison for each class between our algorithm and the work in Ref. 23 is presented in Table 2 using the performance metric given in Eq. (16).

Since there is no text zone in the ground-truth of the *Color-Seg-Images* document type, this class is not considered in the analysis. The gray-scale version of the documents are tested in Ref. 23 although they are scanned in RGB color space. Recall that our system has no dependency on the number of channels of a scanned document. Music and line-drawing documents are excluded in our data-set. In most of the documents, both algorithms provide almost the same accuracy rates. Our technique significantly outperforms against the method in Ref. 23 for article, correspondence, and math document types as shown in Table 2. On the contrary, the study in Ref. 23 gives considerably better classification results in check and program listing. When all the document types are considered, the overall performance of the proposed algorithm achieves 87% accuracy, while the methodology in Ref. 23 performs with an average of 81% accuracy rate.

3.1.2 *Comparison to work done by Won*

An algorithm for extracting images in digital documents has been proposed by Won,[35] where 233 article scanned documents provided by MediaTeam [40] are utilized to evaluate its performance. According to the given ground-truth by the Oulu database of the 233 article documents, 47 documents contain one or more photo regions (document with images), while the rest do not contain any photo region (documents w/o images). The performance results are presented in terms of total error rates, false negatives, and false positives as shown in Fig. 18. Pixel accuracy is used as a performance metric to obtain the rates in Fig. 18, where documents that have photos introduce more false positives in our technique compared to the study in Ref. 35. In our algorithm, the optimal block-size is fixed, and the block-size reduction is not applied. Therefore, the pixels at the boundaries are detected as part of the photo region, although they are nonphoto pixels. However, nonphoto regions are more accurately classified in our system as seen in the false negative plot. Moreover, our false positive rates in documents without photos are less than the method in Ref. 35. This implies that the preprocessing module in our system provides a document that has enhanced text and photo regions for classification purposes. It is worth mentioning that the rate of false negatives in documents that do not contain photo regions is zero. The average error rate for all 233 documents is ~2.9% in Ref. 35 [see Fig. 18(a)] while it is ~4.1% in our study [see Fig. 18(b)]. Note that the technique in Ref. 35 is limited to gray-scale scanned documents only. It also assumes that these documents have a white background region. In contrast, our algorithm makes no such assumptions and achieves 91% photo classification accuracy in ~500 documents for both RGB and gray-scale scanned documents. In addition, the technique in Ref. 35 is evaluated on article type documents only while our algorithm is tested on the entire database [see Table 1].

4 Conclusions

In this paper, a page classification technique has been proposed where text, photo, and strong edge/line regions are detected for both color and gray-scale scanned documents. Our methodology applies wavelet decomposition, RLE, projection based on basis vectors, MRF-MAP optimization, Hough transform, edge linking, and a K-Means merging procedure to differentiate between text, photo, and background data. The performance of the algorithm has been demonstrated on a large database of simple to complex text, photo, and background gray-scale and color scanned documents with an average accuracy of 89% in comparison to ground-truth data. In addition, our study provides consistent results for different types of documents. It differs from prior art by employing a preprocessing stage to "normalize" the

input followed by appropriately designed techniques to locate text, photo, and background information.

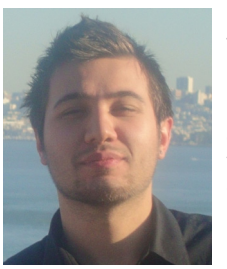
The proposed region classification methodology can be employed in systems for object-oriented printing and OCR applications to name a few. As a future work, extensions to encompass additional objects such as gradients, one-dimensional and 2-D bar-codes are being considered.

Acknowledgments

This research was supported by Hewlett Packard Corporation and the Electrical and Microelectronic Engineering Department of the Rochester Institute of Technology. The authors wish to thank Mr. Sreenath Rao Vantaram, Abdul Haleem Syed and Kuldeep Shah for their valuable comments.

References

1. L. O' Gorman and R. Kasturi, *Document Image Analysis*, IEEE Computer Society Press, Los Alamitos, CA (1997).
2. A. Khandelwal et al., "Text line segmentation for unconstrained handwritten document images using neighborhood connected component analysis," in *Pattern Reg. Mach. Intell, Lec. Notes Comp. Sci.* Vol. 5909, pp. 369–374 (2009).
3. S. Mandal et al., "Detection and segmentation of table of contents and index pages from document images," in *2nd International Conference on Document Image Analysis for Libraries*, IEEE, Lyon, France (2006).
4. S. Yu et al., "Improving pseudo-relevance feedback in web information retrieval using web page segmentation," in *Proc. 12th international conference on World Wide Web*, pp. 11–18, ACM, New York (2003).
5. M. Mitra and B. Chaudhuri, "Information retrieval from documents: a survey," *Info. Retr.* **2**(2), 141–163 (2000).
6. H. Fujisawa, Y. Nakano, and K. Kurino, "Segmentation methods for character recognition: from segmentation to document structure analysis," *Proc. IEEE* **80**(7), 1079–1092 (1992).
7. S. Mao, A. Rosenfeld, and T. Kanungo, "Document structure analysis algorithms: a literature survey," *Proc. SPIE* **5010**, 197–207 (2003).
8. J. Tse et al., "An OCR-independent character segmentation using shortest-path in grayscale document images," in *6th International Conference on Machine Learning and Applications*, Vol. 2007, pp. 142–147, IEEE, Cincinnati, OH (2007).
9. J. Fisher, S. Hinds, and D. D'Amato, "A rule-based system for document image segmentation," in *Proc. 10th International Conference on Pattern Recognition*, Vol. 1, pp. 567–572, IEEE, Atlantic City, NJ (1990).
10. F. Esposito et al., "An experimental page layout recognition system for office document automatic classification: an integrated approach for inductive generalization," in *Proc. 10th International Conference on Pattern Recognition*, Vol. 1, pp. 557–562, IEEE, Atlantic City, NJ (1990).
11. R. Haralick, "Document image understanding: geometric and logical layout," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 385–390, IEEE, Seattle, WA (1994).
12. A. Zlatopolsky, "Automated document segmentation," *Pattern Recogn. Lett.* **15**(7), 699–704 (1994).
13. D. Sharma and B. Kaur, "Document image segmentation using recursive top-down approach and region type identification," *Commun. Comput. Info. Sci.* **70**, 571–576 (2010).
14. Z. Shi and V. Govindaraju, "Multi-scale techniques for document page segmentation," in *Proc. 8th International Conference on ICDAR*, Vol. 2, pp. 1020–1024, IEEE Computer Society, Washington, DC (2005).
15. F. Wahl, K. Wong, and R. Casey, "Block segmentation and text extraction in mixed text/image documents," *Comput. Graph. Image Process.* **20**(4), 375–390 (1982).
16. S. Lam, D. Wang, and S. Srihari, "Reading newspaper text," in *Proc. 10th International Conference on Pattern Recognition*, Vol. 1, pp. 703–705, IEEE, Atlantic City, NJ (1990).
17. A. Antonacopoulos and R. Ritchings, "Flexible page segmentation using the background," in *Proc. 12th International on Pattern Recognition: Computer Vision & Image Processing*, Vol. 2, pp. 339–344, IEEE, Jerusalem, Israel (1994).
18. D. Drivas and A. Amin, "Page segmentation and classification utilising a bottom-up approach," in *Proc. 3rd ICDAR*, Vol. 2, pp. 610–614, IEEE, Montreal, Quebec, Canada (1995).
19. A. Jain and B. Yu, "Document representation and its application to page decomposition," *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(3), 294–308 (1998).
20. A. Simon, J. Pret, and A. Johnson, "A fast algorithm for bottom-up document layout analysis," *IEEE Trans. Pattern Anal. Mach. Intell.* **19**(3), 273–277 (1997).
21. S. Grover, K. Arora, and S. Mitra, "Text extraction from document images using edge information," in *Annual IEEE India Conference (INDICON)*, Vol. 1–4, IEEE, Gujarat, India (2009).
22. A. Jain and S. Bhattacharjee, "Text segmentation using Gabor filters for automatic document processing," *Mach. Vis. Appl.* **5**(3), 169–184 (1992).
23. J. Duong et al., "Extraction of text areas in printed document images," in *Proc. 2001 ACM Symposium on Document engineering*, pp. 157–165, Springer-Verlag, New York (2001).
24. T. Randen and J. Husoy, "Segmentation of text/image documents using texture approaches," in *Proc. the NOBIM-konferansen*, pp. 60–67, NOBIM, Asker, Norway (1994).
25. K. Tombre et al., "Text/graphics separation revisited," in *Proc. 5th International Workshop on Document Analysis Systems V*, pp. 200–211, Springer-Verlag, London (2002).
26. M. Lin, J. Tapamo, and B. Ndovie, "A texture-based method for document segmentation and classification," *South African Comput. J.* **36**, 49–56 (2006).
27. S. Wang and H. Baird, "Feature selection focused within error clusters," in *19th International Conference on Pattern Recognition*, pp. 1–4, IEEE, Tampa, FL (2008).
28. S. Wang, H. Baird, and C. An, "Document content extraction using automatically discovered features," in *10th International Conference on Document Analysis and Recognition*, pp. 1076–1080, IEEE, Barcelona, Spain (2009).
29. S. Chaudhury, M. Jindal, and S. Dutta Roy, "Model-guided segmentation and layout labelling of document images using a hierarchical conditional random field," *Pattern Recogn. Mach. Intell.* **5909**, 375–380 (2009).
30. H. Baird et al., "Document image content inventories," in *Proc. SPIE/IS&T Document Recognition & Retrieval XIV Conference*, Vol. 6500, SPIE, San Jose, CA (2007).
31. Y. Zheng, H. Li, and D. Doermann, "Machine printed text and handwriting identification in noisy document images," *IEEE Trans. Pattern Anal. Mach. Intell.* **26**(3), 337–353 (2004).
32. S. Kumar et al., "Text extraction and document image segmentation using matched wavelets and MRF model," *IEEE Trans. Image Process.* **16**(8), 2117–2128 (2007).
33. C. Shin, D. Doermann, and A. Rosenfeld, "Classification of document pages using structure-based features," *Int. J. Doc. Anal. Recogn.* **3**(4), 232–247 (2001).
34. L. Caponetti, C. Castiello, and P. Górecki, "Document page segmentation using neuro-fuzzy approach," *Appl. Soft Comput.* **8**(1), 118–126 (2008).
35. C. S. Won, "Image extraction in digital documents," *J. Electron. Imag.* **17**(3), 033016 (2008).
36. S. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.* **11**(7), 674–693 (1989).
37. K. Wong, R. Casey, and F. Wahl, "Document analysis system," *Res. Dev.* **26**(6), 647–656 (1982).
38. N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man and Cybernetics, Part A (Systems & Humans)* **9**(1), 62–66 (1979).
39. H. Derin and H. Elliott, "Modeling and segmentation of noisy and textured images using gibbs random fields," *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-9**(1), 39–55 (1987).
40. J. Sauvola and H. Kauniskangas, "Media team document database ii," http://www.mediateam.oulu.fi/downloads/MTDB/, University of Oulu, Finland (1999).



M. Sezer Erkilinc is a MPhil/PhD candidate in the optical networks research group in the Department of Electronics and Electrical Engineering at the University College London in London, UK. He received his BS and MS degree in electrical engineering from Koc University, Istanbul, Turkey, in 2009 and Rochester Institute of Technology (RIT), Rochester, New York in 2011, respectively. His main research interests are in the areas of developing FPGA based real-time DSP techniques for dispersion compensation and advanced modulation formats in optical networks, image understanding, and content-based document analysis. He is a graduate student member of SPIE and IEEE.

Mustafa Jaber is a computer vision scientist at IPPLEX Holdings Corporation, Santa Monica, California. He received his BS in electrical engineering from the Islamic University of Gaza, Gaza, Palestine, in 2003, and MS in the same discipline from the Rochester Institute of Technology (RIT), Rochester, New York, in 2007. He also received his PhD in imaging science from the Chester F. Carlson Center for Imaging Science at RIT in 2011. His research interests are in the areas of

digital image understanding and statistical image processing. He is a member at the IEEE and the SPIE organizations.

Eli Saber is a professor in the Electrical and Microelectronic Engineering Department at the Rochester Institute of Technology. Prior to that, he worked for Xerox Corporation from 1988 until 2004 in a variety of positions ending as product development scientist and manager at the Business Group Operations Unit. He received a BS degree in electrical and computer engineering from the University of Buffalo in 1988, and the MS and PhD degrees in the same discipline from the University of Rochester in 1992 and 1996, respectively. From 1997 until 2004, he was an adjunct faculty member at the Electrical Engineering Department of the Rochester Institute of Technology and at the Electrical & Computer Engineering Department of the University of Rochester. His research interests are in the areas of digital image and video processing including image/video segmentation, object tracking, content-based image/video analysis and summarization, multi-camera surveillance processing, three dimensional scene reconstruction, and image/video understanding. He is a senior member of the Institute for Electrical and Electronic Engineers (IEEE) and the Imaging Science and Technology (IS) society, a member of Image and Multi-dimensional Digital Signal Processing (IMDSP) Technical Committee and member and former chair of IEEE Industry Technology Committee (ITT). He served as an associate editor for the IEEE Transaction on Image Processing and is currently serving as an area editor for the *Journal of Electronic Imaging*. He has served on several conference committees, including finance chair for ICIP 2002 and ESPA 2012, tutorial chair for ICIP 2007 and ICIP 2009, and general chair for ICIP 2012. He holds many conference and journal publications in the area of image and video processing.

Peter Bauer biography and photograph not available.

Dejan Depalov received a BS degree in electrical engineering and computer science from the University of Illinois, Chicago, in 2001, and an MS and PhD in electrical and computer engineering from Northwestern University, Evanston, Illinois, in 2003 and 2007, respectively. He is currently working at Hewlett-Packard as an imaging engineer.