Pattern Recognition 42 (2009) 1419 – 1444

Contents lists available at ScienceDirect

Pattern Recognition

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

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

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

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

ARTICLE INFO

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

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

ABSTRACT

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

1. Introduction

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

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

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

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

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

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

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

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

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

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