SIGRAD 2016
M. Hayashi (Editors)

Low Quality Mobile Image Data Processing Under Uneven Shading

- Separating and Cleaning Text Lines and Graphic Regions in mobile Color Document Image -

Xiaohua Zhang[1], Ning Xie[2], Masayuki Nakajima[3,4], Masaki Hayashi[4] and Steven Bachelder[4]

[1]Hiroshima Institute of Technology, Hiroshima, Japan
[2]Tongji University, Shanghai, China
[3]Kanagawa Insitute of Technology, Atsuki, Japan
[4]Uppsala University, Visby, Gotland, Sweden

Abstract

This paper proposes a simple approach for extracting texts from graphic regions in low quality color document images taken by smart phones or other mobile devices with cameras. An algorithm first computes an edge map by the Canny edge detector. All textual and non-textual regions are then analyzed heuristically based on their connected components(CC). A 2D histogram is calculated to estimate the frequent width and height of connected components. After grouping the CCs according to association rules, the CCs in which the width or height levels are then measured as extremely large or small are assigned as non-textual regions. The remaining CCs are then extracted as text regions. The results of our experimentations demonstrate that the proposed approach performs with plausible consistency.

Categories and Subject Descriptors (according to ACM CCS): I.4.8 [IMAGE PROCESSING AND COMPUTER VISION]: Scene Analysis—Shading

1. Introduction

Smart phones and other mobile devices equipped with camera are becoming increasingly ubiquitous with widespread use throughout most parts of the world. An effective way of taking a memo of text segments found in magazines, advertisements or other texts is to simply take a photo of the text rather than resorting to pen and paper. To increase automatisation, the text in the document image should be formatted in order to facilitate text recognition [ZENM13]. In this paper, we propose a simple approach for extracting texts from graphic regions in low quality color images. The algorithm computes an edge map using the Canny edge detector. By analyzing connected components heuristically, the textual and non-textual regions are separated making text extraction possible.

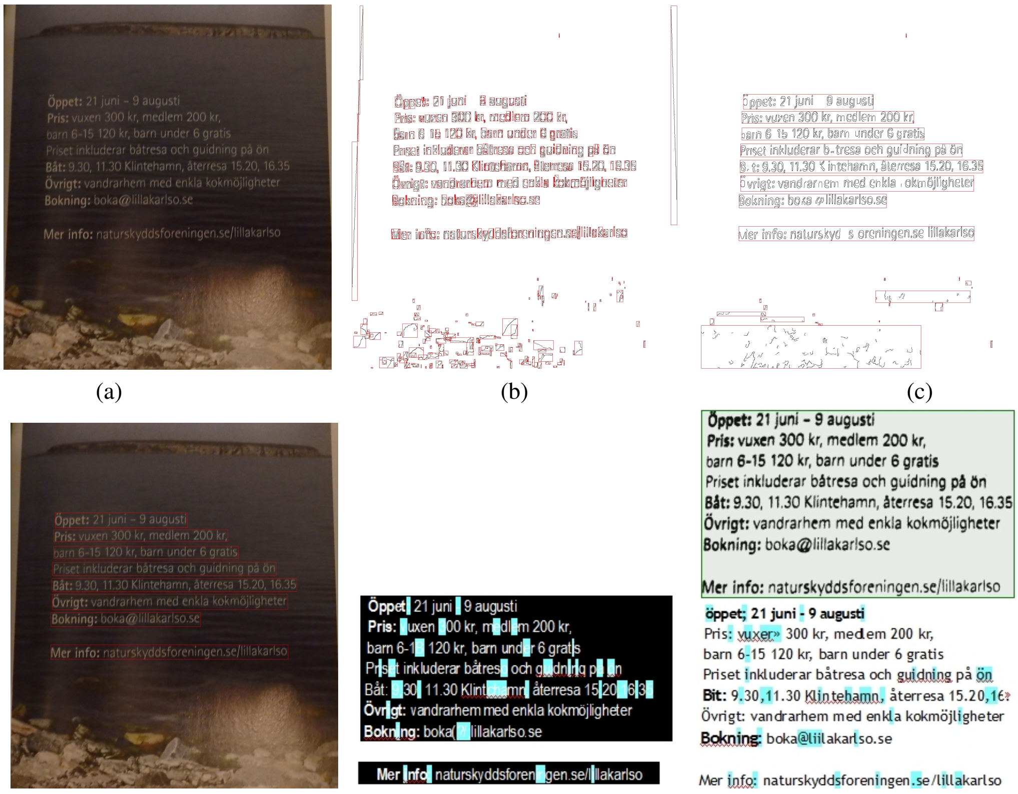
2. Extracting text regions

In a document image such as shown in Fig. 1(a) which is taken by a mobile phone under poor lighting conditions, the image contains textual and non-textual regions and is covered with uneven shading. We assume that the given document image has little distortion and that the text lines are nearly aligned in horizontal.

The proposed algorithm is applied to the gray channel. Once the uneven shading is removed [ZXHX16] based on retinex theory [LM71], an edge map is constructed using the Canny algorithm. Note that since the background in the image is complicated, the binarized is useless for text extraction. The detected edges are very important and are used for extracting text lines from graphic regions.

To separate textual and non-textual regions, all pixels on edges are labeled. After labeling, each connected component (CC) is embraced as a quadrangle called a bounding box as shown in Fig. 1(b). A CC is featured with several attributes such as position, width, height, area, density, aspect ratio etc. If a CC has a large width or height, such as those on the left and right in Fig. 1(b), it is obviously a non-textual region and should be deleted. The CCs with small area, density or aspect ratio are considered as noise and are deemed non-textual regions. These CCs are also deleted.

A text line consists of several CCs distributed on the same line. To group these CCs, the frequent width and height of


Öppet: 21 juni - 9 augusti
Pris: vuxen 300 kr, medlem 200 kr,
barn 6-15 120 kr, barn under 6 gratis
Priset inkluderar båtresa och guidning på ön
Båt: 9.30, 11.30 Klintehamn, återresa 15.20, 16.35
Övrigt: vandrarhem med enkla kokmöjligheter
Bokning: boka@lillakarlso.se
Mer info: naturskyddsforeningen.se/lillakarlso

(a) (b) (c)

(d) (e) (f)

Figure 1: *Original low quality color document image (a); initial connected components after the Canny edge detection (b); connected components after grouping (c); extracted text lines (d); OCR result of text regions (e)*

CCs are required. To this end, a 2D histogram of width and height is computed. The location of the highest peak indicates the mode of width and height. Compared with the frequent height, if a CC has a larger height, it is assumed as an image region. All inner CCs are combined with its mother CC by detecting the inclusion relations. Later, all CCs are grouped according to their attributes to extract text lines as demonstrated in Fig. 1 (c). After grouping, connected components with extremely large or small width and height as those at the bottom in Fig. 1 (c) are deleted.

All remaining connected components are the extracted text lines as shown in Fig. 1 (d). It can be clearly observed that the text lines are well extracted. We binarized [SP00] the extracted text regions as shown at the top in Fig. 1 (f) and proceeded to feed it to OCR software, the recognized result is at the bottom in Fig. 1 (f). The recognition rate is very high. Note that the recognition rate, as shown in Fig. 1 (f), is very low when feeding original image directly to the OCR software.

3. Conclusions

We proposed an approach for extracting text regions from graphic regions in low quality document images taken by smart phones and other mobile devices equipped with cameras. Once the uneven shading was removed, the text lines were collected by heuristically analyzing the connected components computed from the detected edges. The text regions were segmented and defined according to the frequent size, positions, densities from all the CCs. To verify the accuracy of extractions, the text-only regions are binarized, then fed to OCR software. The experimental results demonstrate that the recognition rate is higher than images tested without text region extraction.

References

[LM71] Land E., McCann J.: Lightness and retinex theory,. *Journal of the Optical Society of America 61* (1971), 1–11. 1

[SP00] Sauvola J., Pietikainen M.: Adaptive document image binarization. *Pattern Recognition 33*, 2 (Feb. 2000), 225–236. 2

[ZENM13] Zirari F., Ennaji A., Nicolas S., Mammass D.: A document image segmentation system using analysis of connected components. In *Proc. of* 12^{th} *ICDAR* (2013). 1

[ZXHX16] Zhang X., Xie N., Huang H., Xin Y.: Fast restoration of text strokes from a document image with uneven shading. In *Proc. of* 19^{th} *IWAIT* (2016). 1