DESCRIPTION OF THE METHOD

Name: MHS-2019 system

Author: Tuan Anh Tran

Participant: Tuan Anh Tran[a], Nam Quan Nguyen[a], Quoc Thang Nguyen[a], Hai Duong Nguyen[b], Soo Hyung Kim[c].

Affiliation:

(a): HoChiMinh National University City - University of Technology, Viet Nam & Cinnamon AI

(b): Department of Computer Science and Software Engineering, Concordia University, Canada

(c): Department of Computer Science, Chonnam National University, Gwangju, Republic of Korea.

Our method includes principles of operation and steps:

1. Negative/positive image detection

Based on the analysis of background (distribution of dark/white pixel), we identify the input image is a negative or positive image.

2. Binarization

If the input is a positive image, we use the combination of Sauvola and Otsu's method, otherwise, we apply the Otsu method and then invert them.

3. Text and Non-text classification

The main stage of text and non-text classification in the MHS-2019 system is the Minimum Homogeneity Algorithm (MHA) which was first introduced in 2016 [2]. This algorithm based on the connected component analysis [4] in a statistical approach. In 2017, an essential update in the core of this algorithm, the MLL classification [1] which uses the combination of Multilevel and Multilayer Homogeneity structure is presented. In recent years, we develop some module languages for the core of MHS such as Korean, France, Vietnamese, and Japanese.

4. Text segmentation and Image classification.

In this step, text documents are segmented to get text regions, and non-text elements are classified into different types.

4.1. Text segmentation

We apply the combination of text line extraction, paragraph segmentation, and adaptive mathematic morphology to get the text region [1]. It should be noted that the segmentation contains two phases. The second phase will be performed in region refinement step.

4.2. Non-text classification

Based on the properties of non-text elements, we classify them into the negative-text region, line, table, separator, chart, and image [1]. Our system also contains a robust table detection method which was introduced by Tran et al. [3] in 2016.

5. Region refinement and labeling.

Based on the boundary of each region, we extract the rectangular shape of the text and non-text region. In the MHS-2019 we add one more segmentation step in the region which was obtained from step 4.1. It will help us to correct some segmentation errors such as split and miss-classification errors (because at this time we look only to the local region instead of all global regions).

All of the identified regions are labeled (heading, page number, etc.) based on its text size and position.

6. Optical Character Recognition

All text regions are then recognized via Tesseract OCR in Computer Vision System Toolbox™ (Matlab).

Reference:

[1] T. A. Tran, I. S. Na, S. H. Kim, "A Robust System for Document Layout Analysis using Multilevel Homogeneity Structure," Expert Systems With Applications, vol. 85, pp. 99-113, 2017.

[2] T. A. Tran, I. S. Na, S. H. Kim, "Page Segmentation using Minimum Homogeneity Algorithm and Adaptive Mathematical Morphology," International Journal on Document Analysis and Recognition, vol. 19, pp. 191-209, 2016.

[3] T. A. Tran, H. T. Tran, I. S. Na, S. H. Kim, G. S. Lee, H. J. Yang, "A Mixture Model using Random Rotation Bounding Box to Detect Table Region in Document Image," International Journal of Visual Communication and Image Representation, vol. 39, pp. 196-208, 2016.

[4] T. A. Tran, I. S. Na, S. H. Kim, "Separation of Text and Non-text in Document Layout Analysis using a Recursive Filter," KSII Transaction on Internet and Information Systems, vol. 9, pp. 4072-4091, 2015.