Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 93 (2016) 469 – 477

Procedia Computer Science

6th International Conference On Advances In Computing & Communications, ICACC 2016, 6-8 September 2016, Cochin, India

Text/Image Region Separation for Document Layout Detection of Old Document Images using Non-linear Diffusion and Level Set

Sachin Kumar S[a]*, Parvathy Rajendran[a], Prabaharan P[b], K P Soman[a]

[a]*Centre for Computational Engineering and Networking (CEN), Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, Coimbatore, India*

[b]*Amrita Centre for Cyber Security and Networks, Amrita School of Engineering, Kollam, Amrita Vishwa Vidyapeetham*

Abstract

Text/Image region separation is the process of identifying location of various text and image regions in a scanned document image. This is particularly helpful in detecting the layout of a scanned document image. The text region thus obtained can be used for optical character recognition (OCR) operation. The text region can be used to label and train automatic layout learning system to detect locations of title, keywords, subheadings, paragraphs, image locations etc. In case of regular image and text boundaries, Profiling or morphological operations can be used for separating the text and image regions and to achieve correct document layout out detection. However, the real-world documents will have irregular boundaries and noise, the usual profile based methods and its heuristic often fails. This will lead to incorrect document layouts. This paper proposes to use edge enhancement diffusion and level set method for text/image region separation from scanned document images. The result obtained shows that the proposed method works when the document contain multiple images. The proposed method detects the layout of the scanned document even when the image and the text regions have irregular shape.

Peer-review under responsibility of the Organizing Committee of ICACC 2016

Keywords: document layout, diffusion, level set, edge enhancement, edge preserving

1. Introduction

Digital copy of documents provides the flexibility of storing physical documents in a more organized and makes it easily manageable/searchable. Such digital archives preserves the documents for longer periods. Digital copies serves the purpose for preserving the documents or to make the document available online. The physical files for

* Corresponding author. Tel.: +91-9500823956
E-mail address:sachinnme@gmail.com

Peer-review under responsibility of the Organizing Committee of ICACC 2016
doi:10.1016/j.procs.2016.07.235

long periods may get destroyed, eroded or distorted etc. Making digital copies of old documents sometimes preserves the heritage value associated with it. While several documents are digitally archived, few scanned documents of old story book known as "Chandamama, Ambilimama etc", are significant and stunning as it reflects the tradition of Indian culture, taking stories from medieval period. The characters and subject of the stories touched all the corners of the society. Those magazines had a unique style of story-telling, mostly like grandparents style of story-telling (a third-person narration). It finds it place way before the comic magazines and graphics-rich novels. These magazines had once readers falling in different age groups, from young to grown-ups. The stories and folktales were enriched good subjects in order to captivate the minds and guide the generations by giving the view about life, moral values of life, helping to discard bad and accept good things etc. Those old magazines were truly rich with edutainment. However, the advent and growth of entertainment industry and Television media, decades after, those popular magazines witnessed its fall out and the physical copies are few in numbers or not available. In order to preserve those precious contents and to deliver it to the present young generation, the available old documents need to be preserved and enhanced digitally. The enhancement of the scanned document images of the old physical copy is an inherent procedure as it would had undergone distortions, degradations or noises etc. This paper presents an approach using non-linear diffusion and level set method to separate the text and image regions from the old scanned document images. The approach also understands the layout of the document to detect the title, words, lines, tables, paragraph. In a document, locating the text region is important as it carries majority of the information regarding the document. The image region mostly acts as a supporting material to the text content. The extracted text region can be used for OCR operation. Converting the scanned document into text form and compressed image reduces the storage space than the original one. The text/image region separation inherently tries to find out the layout of the document or does support layout detection. The segmentation operation and layout analysis of the scanned document are divided into bottom-up, top-down, and hybrid[1,2,3]. A scanned document can get affected with dust or spots etc during the scanning and this introduces noise. Other ways of getting noise will be due to ageing, degeneration, photocopying etc. Through image enhancements or filtering operations, noise can be reduced or removed. Smoothing or blurring operations also reduces the noise. Blurring is used in pre-processing stage as it can remove the smaller unwanted details from an image. In the binary document of degraded image, smoothing fills the little gaps between characters to correct its edges. Filters on the other hand provides a value for a pixel using its neighbourhood pixel values. Several PDE based methods such as ROF model, Sobolev model, L1-norm regularization method etc are also popular for noise reduction[4]. Layout analysis can be broadly divided as physical and logical or functional layout analysis. Physical layout refers to the several boundary regions in the scanned document. These regions can be title, words, lines, tables, paragraph, symbols, images, log etc. Such entities of the regions forms the functional aspect of the layout analysis. These can be used to find the inter-relationship between the entities helps to find the logical relation in between. These relationship can give information regarding the logical construct of the regions. The general approach for layout analysis are bottom-up and top-down. In bottom-up approach, algorithm gather pixels and repeatedly group to form regions such as words, lines, paragraphs. Whereas in top-down approach, the algorithm starts from the scanned document image and repeatedly divides into smaller regions. It stops the activity when a certain defines condition is satisfied. The hybrid approach utilizes both the approaches.

Several article discuses the layout analysis problem in scanned document images. The main purpose of the document layout analysis identifoies different physical regions in a document. The X-Y Cut algorithm for partitioning document was discussed in[5]. It does it by projecting the regions alternatively on the vertical and horizontal axis. In[6], smearing algorithm known as run length algorithm is used. It does binarization by smearing the black pixels. The small white pixels are made black in this algorithm. The paper[7] discuses connected component approach for to group regions. It does this by taking an initial measurement from the document image. [8] discusses an algorithm for segmenting document image by detecting the white spaces in between the pixel columns. In doing this it makes a white rectangle and it stops when it can't make rectangles without the presence of black pixels. The rectangles formed in this manner was considered as the regions of layout. The paper presented in[9] discusses two-steps to find lines of texts. As the initial step it detects the maximal whitespace rectangles which represents the pixel columns. In the second step, these whitespace rectangles becomes an input to the least-square based line detection algorithm. This method tried to find the text-lines. In[10], it discusses a bottom-up algorithm which initially gets few points from the connected component boundaries. Using these points, the algorithm makes a Voronoi diagram. Using a threshold Superfluous Voronoi edges are removed. The algorithm uses a threshold for height, width and aspect-ratio. The paper[11] generates rectangular regions by using a modified Manhatton layout. In[12], uses a top-down method for layout analysis. The document images are segmented initially and later everything is grouped in a