Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 45 (2015) 205 – 214

Procedia Computer Science

International Conference on Advanced Computing Technologies and Applications (ICACTA-2015)

Automatic Removal of Handwritten Annotations from Between-Text-Lines and Inside-Text-Line Regions of a Printed Text Document

P. Nagabhushan, Rachida Hannane, Abdessamad Elboushaki, Mohammed Javed*

Department of Studies in Computer Science, University of Mysore, Mysore-570006, India

Abstract

Recovering the original printed text document from handwritten annotations, and making it machine readable is still one of the challenging problems in document image analysis, especially when the original document is unavailable. Therefore, our overall aim of this research is to detect and remove any handwritten annotations that may appear in any part of the document, without causing any loss of original printed information. In this paper, we propose two novel methods to remove handwritten annotations that are specifically located in between-text-lines and inside-text-line regions. To remove between-text-line annotations, a two stage algorithm is proposed, which detects the base line of the printed text lines using the analysis of connected components and removes the annotations with the help of statistically computed distance between the text line regions. On the other hand, to remove the inside-text-line annotations, a novel idea of distinguishing between handwritten annotations and machine printed text is proposed, which involves the extraction of three features for the connected components merged at word level from every detected printed text line. As a first distinguishing feature, we compute the density distribution using vertical projection profile; then in the subsequent step, we compute the number of large vertical edges and the major vertical edge as the second and third distinguishing features employing Prewitt edge detection technique. The proposed method is experimented with a dataset of 170 documents having complex handwritten annotations, which results in an overall accuracy of 93.49% in removing handwritten annotations and an accuracy of 96.22% in recovering the original printed text document.

Peer-review under responsibility of scientific committee of International Conference on Advanced Computing Technologies and Applications (ICACTA-2015).

Keywords: Handwritten annotation removal; Marginal annotation removal; Between-text-line annotations; Inside-text-line annotation.

* Corresponding author: Tel: +91-97-4116-1929

E-mail address: javedsolutions[at]gmail.com; mohammed.javed.2013[at]ieee.org

Peer-review under responsibility of scientific committee of International Conference on Advanced Computing Technologies and Applications (ICACTA-2015).

doi:10.1016/j.procs.2015.03.123

1. Introduction

Annotating a printed text document refers to the process or the act of adding annotations or writing critical commentary or explanatory notes into the machine printed text documents. Adding annotations at different regions of the printed text document is the habit of many people whenever they read any document. These annotations may be some important observations or corrections marked in their own style; this makes the problem of recovering the original text document very challenging, particularly when the original document is not available.

A reader can add different types of annotations in any part of the document; the type of annotations depend on the information being read by the reader. For example, the annotations can be lines underscoring or side scoring to highlight keywords, sentences or part of a paragraph and so on, or even enclosing such a part with a circular, elliptical or contour shapes holding a word or a sentence of interest, or it can be simply a question mark (when the reader does not understand the content), also it can be in the form of more frequent annotations that are handwritten comments (in which the related idea is extrapolated or the missed details are added), or it can be any other type of external remark that can be added anywhere in a document including: in marginal area of the document, in between-text-line regions, inside-text-line regions crossing over the printed text lines. Based on the regions where annotations appear, we classify them into the following categories: marginal annotations, between-text-line annotations, inside-text-line annotations, and overlapping annotations (see Fig. 1).

There are a few attempts reported in the literature related to handwritten annotations in printed text documents[2,4,5,7]. However, we understood that the nature of the annotations considered in[2,4,5,7] is predefined in structure and location. The method described in[5] by Mori and Bunke, have achieved high extraction rates by limiting the colors or types of annotations. Although the correction is successful, their method has the limitation on colors of annotations as well as type of documents. Their method does not focus on the extraction of annotations but on the utilization of extracted annotations. Guo and Ma[4] proposed a method that involves the separation of handwritten annotations from machine printed text within a document. Their algorithm is based on the theory of hidden Markov models to distinguish between machine-printed and handwritten materials, and the classification is performed at the word level. Handwritten annotations are not limited to marginal areas and also their approach can deal with document images having handwritten annotations overlaid on machine-printed text. However, their extracted annotations are limited to some predefined characters; their method cannot extract handwritten line drawings which is also a frequently used annotation. Y. Zheng et al.[7] proposed a method to segment and identify handwritten text from machine printed text in a noisy document; their novelty is that they treat noise as a distinguished class and model noise based on selected features. Trained Fisher classifiers are used to identify machine printed text and handwritten text from noise. Lincoln Faria da Silva et al.[2] proposed a method that involves the recognition of the handwritten text and machine printed text in a scanned document image. New features for classification of the handwritten text and machine printed text are proposed as well: Vertical Projection Variance, Major Horizontal Projection Difference, Pixels Distribution, Vertical Edges, and Major Vertical Edge. Unfortunately, their method can work only when the machine printed text and handwritten texts are separated. However, in our proposed method, we explore the feature of vertical Prewitt edge in order to remove the handwritten annotations inside the printed text line regions.

To our best knowledge, the proposed methods in the literature do not present a systematic approach for handling the different annotations that appear in printed text documents. Therefore, in order to systematically handle the different annotations coined in this paper, we propose a Handwritten Annotation Removal System (HARS) based on the different types of annotations that a document can have (see Fig. 2). However in this paper, we specifically focus on detecting and removing annotations located in between text lines and inside text line regions. Rest of the paper is organized as follows: In section 2, we detail the proposed methods to remove between-text-line and inside-text-line annotations. In section 3, the experimental results of our proposed methods are reported, and the last section 4 concludes the paper.

2. Proposed Methods

The overall proposed system (HARS) for detecting and removing handwritten annotations consists of five major stages (see Fig. 2), where the scanned annotated document is taken as an input and converted to a binary image; this