A typed and handwritten text block segmentation system for heterogeneous and complex documents (2012)

A Resume

November 30, 2018

Abstract

A method for handwritten text segmentatin, based on feature extraction and learning.

1 Introduction

proposing a connected component oriented approach for text identification and segmentation. The particularity of the approach relies on the fact that the system can handle heterogeneous and complex documents thanks to a learning based approach.

2 Method

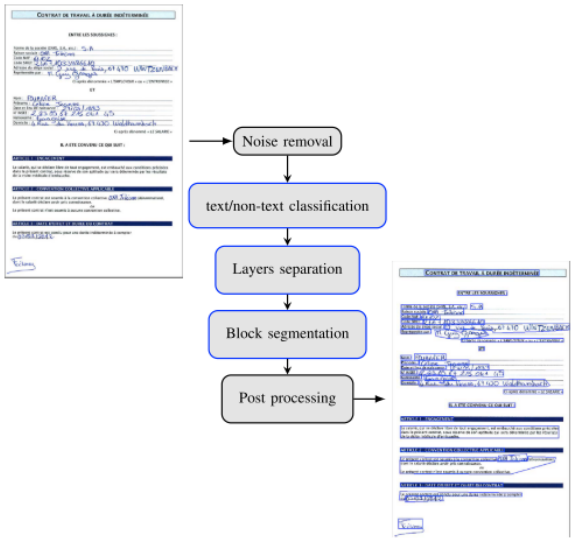

Figure 1: Overview of the proposed system

The system is composed of several detectors (text, tables, images...) that work in parallel. The different detectors work at different levels of the document: image of the document, connected components, lines, and blocks, depending on the nature of the objects to be detected.

The preprocessing consists in filtering small connected components (CC) as well as large CC's close to the borders of the document. The first main step is a CC text/non-text classification. It

consists in extracting simple shape features to classify CC's into text or non-text components. The second main step is a layer separation that consists in separating textual CC's into typed components and handwritten components. Finally, the third step consists in a block segmentation based on the search of empty rectangles applied on the three layers (non-text, typed and handwritten) previously obtained. Finally, a post processing stage combines blocks between handwritten and typed layers in order to reduce segmentation errors by removing small handwritten blocks included in a typed block and vice versa.

3 Text/Non-Text Detection

A key step in our system is the discrimination of each connected component into text or non-text component. by extracting simple features representing the shape of the connected components and its neighborhood, and feed it to MLP classifier.

Feature extraction for each connected component, a set of simple features is extracted: aspect ratio, area ratio, density, compactness, eccentricity, number of connected components included in the current connected component, and number of connected components overlapped with the current connected component.

Classification A MLP is trained on a set of 2000 document images from the MAURDOR training dataset containing both text and graphic components.

4 Layers seperation

The classification between handwritten and typed components relies on a codebook based approach, the learning stage relies on two steps. First, a codebook is built. It contains a collection of contour fragments extracted from a first connected components learning dataset. Then, a MLP classifier is learnt using as features the histogram of occurrences of the fragments of this codebook in the connected components of a second learning dataset.

Codebook construction 1- Fragment extraction and representation: Extract fragments of external contour of connected components. 2- Codebook generation: The codebook generation step aims at finding a collection of similar contour fragments in a first learning dataset with four languages (Ar, Eng, Latin, Fr) to discriminate between types and handwritten text with different languages. A MLP is thus trained on this dataset containing approximately 25000 samples of each class (Arabic typed, Latin typed, Arabic handwritten and Latin handwritten).

5 Block segmentation

This section describes the approach for producing the homogeneous areas using the connected components and their classification. First an aggregating method is used to group the connected components into bigger entities using RLSA; and then a detection of vertical and horizontal white spaces allows to produce a mask that segments the document into areas.

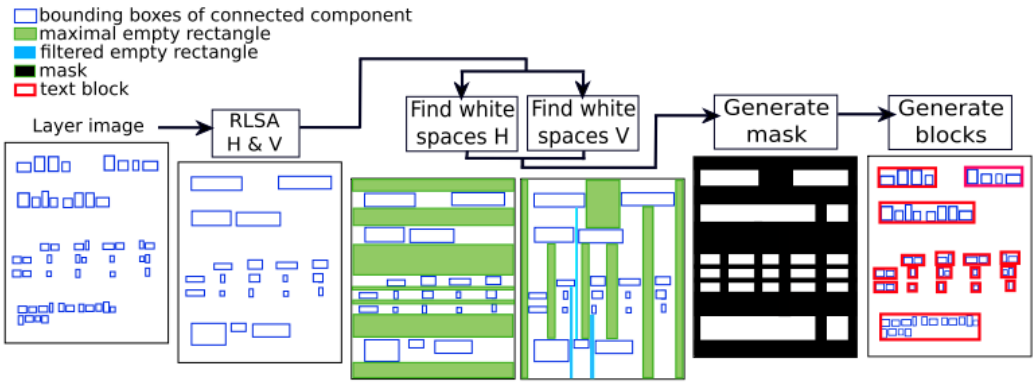

Figure 2: Block segmentation Aggregating

References

[1] P. Barlas, S. Adam, C. Chatelain, T .Paquet, *A typed and handwritten text block segmentation system for heterogeneous and complex documents*