Pattern Recognition 48 (2015) 1197–1210

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

A flexible framework for online document segmentation by pairwise stroke distance learning

Adrien Delaye *, Kibok Lee

Samsung Electronics, DMC R&D Center - MultiMedia R&D Team - Display Processing Lab, Republic of Korea

ARTICLE INFO

Article history:
Received 24 May 2014
Received in revised form
17 September 2014
Accepted 16 October 2014
Available online 27 October 2014

Keywords:
Document analysis and recognition
Online handwriting
Segmentation
Layout analysis
Clustering
Pen computing

ABSTRACT

A flexible framework for handling a variety of segmentation problems in online handwritten documents is introduced. The strategy relies on single linkage clustering and a pairwise stroke distance that is globally trained for direct optimization of the segmentation. We define a variety of features that can contribute to the pairwise distance definition and show how to select a good combination of features for dealing with a new online handwritten document segmentation task. Our experiments demonstrate the validity of the method on a large range of segmentation tasks over several types of documents from various publicly available databases.

1. Introduction

How to split online ink documents into meaningful objects is a crucial problem for high-level document processing tasks such as recognition, beautification or indexation. As of today, a large number of methods were proposed for solving application-dependent segmentation tasks, but there is a lack of general method that could be applied to a large range of problems, or that could serve as a universal evaluation baseline. General methods free of domain-dependent constraints are especially needed for analyzing unconstrained documents (such as the ones resulting from free note-taking), that can exhibit heterogeneous contents, such as text blocks, math equations, tables, diagrams and drawings [1].

In this work, we introduce a flexible segmentation method based on the Single Linkage Agglomerative Clustering equipped with a properly trained distance function. The clustering parameters (distance and threshold) are trained globally by applying an improved version of the distance training algorithm recently introduced by Yin et al. [2]. The system does not require external knowledge other than the selection of appropriate features when applied to a new segmentation problem. It offers a unique flexibility that makes it suitable for a wide range of segmentation tasks from various types of documents (including heterogeneous documents) and at different levels of granularity. The efficiency of the method is experimentally validated on different segmentation tasks: high-level objects and text lines from heterogeneous unconstrained online documents, shapes and labels from sketched diagrams, and symbols from handwritten mathematical expressions.

Section 2 reviews prior works related to the segmentation of online handwritten documents. In Section 3 we present how to train a pairwise distance function optimized for a target online document segmentation task, based on the Single Linkage Agglomerative Clustering. Different categories of features to be considered for inclusion in the distance definition are proposed in Section 4. Section 5 introduces useful metrics for evaluating the predicted segmentation of online documents, and Section 6 presents the results of our experiments on four different segmentation tasks across five publicly available online documents data sets. Section 7 concludes this study and presents perspectives for future research.

2. Related works

A large number of approaches were proposed for various online document segmentation problems. With the predominance of text-related applications, a majority of segmentation methods are actually dedicated to the structuration of ink into text lines.

2.1. Text line segmentation in online documents

Segmentation of text lines from online documents has received a lot of attention in the past. A method for extracting text lines from

* Corresponding author.
E-mail address: ad.delaye@samsung.com (A. Delaye).

http://dx.doi.org/10.1016/j.patcog.2014.10.022
0031-3203/

heterogeneous online documents (i.e. containing non-text ink) was presented by Zhou et al. [3]. It considers temporal and spatial dimensions of the online strokes through several splitting and merging steps. Strokes are first classified as text or non-text with a Markov Random Field encoding their spatial dependencies, and then text strokes are grouped into text lines based on their distance. Each text line string is further segmented by optimization of a cost function with trained parameters. Then, a temporal merging step consisting of a discriminative classifier performs merge/non-merge decisions, before a final spatial merging step handles the delayed strokes. The method is able to correctly segment more than 90% of text lines in free-form Japanese note taking documents [3]. The approach of Liwicki et al. [4] is very similar but does not handle non-textual strokes. Both approaches make assumptions on the properties of the text lines under the form of an appropriate cost function. For example, assumption of text line linearity, penalization of discrepancy in the heights of text lines, spatial compactness of the text lines are used in [3], penalization of the vertical variance of strokes within a line, prevention of crossing between text lines are used in [4].

In the works of Shilman et al. [5], successive grouping steps based on dynamic programming and heuristic temporal and spatial criteria are applied for structuring online ink into words, lines and paragraphs. In [6], the text line segmentation is formulated as a stroke partitioning problem. A cost function combining text line likelihood, configuration consistency and model complexity is optimized globally by gradient descent. In [7], Blanchard and Artieres introduced a system based on Probabilistic Feature Grammars that are trained for the task of text line detection. Even if it shows great robustness with respect to poorly structured documents, the rule-based nature of the system is a limitation for handling free-form documents that may contain unpredictable structures or types of content.

2.2. Segmentation of non-text documents

If the literature about non-textual online document analysis is abundant [8–10], most recognition approaches extensively rely on domain-specific segmentation methods which are difficult or impossible to generalize to other domains. In application to documents such as flowchart diagrams [9,11,12], technical sketches [10], or mathematical expressions [13], over-segmentation strategies are often applied to generate a large number of segmentation hypotheses based on the input strokes. Hypotheses are then evaluated and combined with the help of domain-specific recognition systems or other form of knowledge (e.g. spatial relationships [14]). The over-segmentation method usually encodes constraints relevant to the domain at hand, for example imposing restrictions on the number of strokes per segment [13,15], on the spatial and temporal overlapping between segments [13,16], etc. These segmentation methods are efficient when the document domain is known and homogeneous, but cannot be applied for general online document segmentation problems, for example in heterogeneous documents, or in the case where no recognition engines or combination rules are available.

2.3. General segmentation in online documents

The general problem of online document segmentation arises when no constraints are imposed to the user, who is then free to compose heterogeneous or poorly structured documents. Since the type of content present in the document is unknown at the time of analysis, more flexible segmentation methods have to be defined. These methods need to be free of assumptions regarding the objects properties and should not rely on a recognition engine.

A first attempt towards general segmentation of heterogeneous documents was made by Jain et al. [17]. They proposed a method based on the Minimum Spanning Tree of the ink strokes previously classified as non-text strokes. On one hand, the method for building and cutting the MST in [17] is simplistic as no training scheme is proposed to define the stroke distance and the cutting criterion is hand tuned. The system is also error-prone because initial text/non-text stroke classification is performed locally and classification errors cannot be corrected in the segmentation step. On the other hand, it offers a great flexibility because it is very simple. Indeed, it boils down to cutting some edges of the MST graph, thus only considering pairwise stroke relationships without assuming properties of the segmented blocks. With respect to applicability in different types of documents or in heterogeneous documents, this is a considerable advantage over methods that apply combinatorial algorithms to optimize cost functions based on expected block features such as the text line segmentation methods discussed above.

Interestingly, other works have exploited the idea of cutting a MST structure for segmentation of documents. For example, in [18], the MST of connected components is built and split at specific edges to retrieve text lines in offline handwritten documents. A learning scheme is adopted to tune the distance metric for this task, and the method shows significant improvements when compared to a manually defined distance. However, although the tree is built based on the trained distance, the criterion for tree segmentation involves block-level properties. Consequently, the distance training task is not directly correlated to the segmentation result and the extension of this method to handling other types of blocks or heterogeneous documents is not straightforward.

In their work [19], Stahovich et al. presented a system for grouping strokes into objects by considering pairwise stroke relationships only. Discriminative classifiers are trained for labeling pairs of strokes as *NearJoin*, *FarJoin* or *NoJoin* based on features describing their spatial and temporal relationships. Experiments show a good segmentation quality for shapes from various sketch domains, but the system is not applied to other types of documents or to heterogeneous documents. Like [17], it also involves a prior stroke classification step that is likely to introduce errors in the final segmentation.

Recently, we have presented systems for labeling strokes and segmenting objects from heterogeneous online documents [1,20,21]. This line of work aims at applying to document analysis some recent progress in computer vision permitted by the use of hierarchical graphical models [22–24]: the combination of several initial representations of an image obtained from different naïve segmentation methods (or different parameters) greatly benefits the prediction of labels and boundaries of objects. But while a large number of methods are available for estimating segmentation of objects in images, there is a lack of equivalent tools for online handwritten documents.

In this work, a new solution is presented to deal with segmentation of online documents by training pairwise stroke distance and applying Single Linkage Agglomerative Clustering. By only considering pairwise stroke relationships, the method is free of constraints regarding blocks properties (as [17–19]), which is an advantage for handling highly variable objects. Unlike in works of Yin et al. [18] and of Stahovich et al. [19], we train the clustering parameters by optimizing an objective function that directly describes the quality of segmentation obtained with the clustering method, following the idea of Yin et al. [2] and improving on their algorithm. The ability of our method to deal with difficult segmentation situations is demonstrated on unconstrained handwritten documents with heterogeneous content, where our system outperforms previously published result on the task of object segmentation [1]. We further experiment the method on several other segmentation tasks, including for the segmentation of text lines from unconstrained documents, of basic shapes from sketched diagrams, and of individual symbols from math equations.