Pattern Recognition 47 (2014) 959–968

Contents lists available at ScienceDirect

Pattern Recognition

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

Contextual text/non-text stroke classification in online handwritten notes with conditional random fields

Adrien Delaye*, Cheng-Lin Liu

National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun East Road, Beijing 100190, PR China

ARTICLE INFO

Available online 2 May 2013

Keywords:
Online handwritten documents
Structured prediction
Stroke classification
Conditional random fields
Text/non-text separation

ABSTRACT

Analysing online handwritten notes is a challenging problem because of the content heterogeneity and the lack of prior knowledge, as users are free to compose documents that mix text, drawings, tables or diagrams. The task of separating text from non-text strokes is of crucial importance towards automated interpretation and indexing of these documents, but solving this problem requires a careful modelling of contextual information, such as the spatial and temporal relationships between strokes. In this work, we present a comprehensive study of contextual information modelling for text/non-text stroke classification in online handwritten documents. Formulating the problem with a conditional random field permits to integrate and combine multiple sources of context, such as several types of spatial and temporal interactions. Experimental results on a publicly available database of freely hand-drawn documents demonstrate the superiority of our approach and the benefit of contextual information combination for solving text/non-text classification.

1. Introduction

Automatic interpretation of free form online handwritten documents is considered as a very challenging task due to the high diversity of contents and the lack of prior knowledge available. By fully exploiting possibilities of pen-based interfaces, one is able to input rich contents, from writing text lines and paragraphs to composing tables, sketching diagrams, realizing free drawings, or gesturing annotations and commands. In a realistic note-taking scenario, no constraint is enforced to the user who has the liberty to create a document without complying with any specific composition rule. As more and more of such rich, heterogeneous documents are acquired from digital pens, pen-enabled computers, smart phones, tablets and electronic whiteboards, there is a need for better analysis and recognition algorithms.

Without loss of generality, it is often assumed that ink documents have some textual content that is of special importance for their interpretation. Contrary to non-textual elements that can have highly variable properties, textual elements present regularities and robust features, such as a hierarchical organisation into words, lines and paragraphs, or a locally stable size. Moreover, textual elements usually convey semantically rich information and they can be analysed by efficient handwriting recognition engines [1], whereas it is difficult to define general analysers for non-textual elements. For these reasons, the task of accurately separating textual from non-textual strokes in an online document is regarded as a crucial step towards general free-form document interpretation [2].

In this paper, we present our advances for the text/non-text classification problem in online handwritten documents with conditional random fields. Contextual interactions between elements of the document are of crucial importance and the use of CRFs permits to model these complex dependencies under a well principled framework. Though many works have been conducted for stroke classification using various features and classifiers, and some considered the interactions between strokes, a thorough exploitation of contextual information was not reported. We conduct a comprehensive study on the exploitation of contextual information for supporting the stroke classification task. To compensate for the lack of prior knowledge, the rich information contained in the online documents is fully exploited: visual information from the spatial distribution of the strokes, and dynamic information from the temporal sequence of writing. Our results establish new state-of-the-art performance for text/non-text stroke classification in free form documents.

The first section of this paper reviews past research works about the separation of textual and non-textual content in online documents. This survey highlights the need for a global study on the exploitation of context and combination of various types of

* Corresponding author. Tel.: +86 15811394302.
E-mail addresses: adrien.delaye@nlpr.ia.ac.cn, adrien.delaye@gmail.com (A. Delaye), liucl@nlpr.ia.ac.cn (C.-L. Liu).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.04.017

interactions for this task. In Section 3, we present the conditional random field model that allows a large flexibility in the exploitation of contextual knowledge. Different definitions for stroke interaction systems are then presented in Section 4, suggesting that complementary neighbourhood systems can be combined in the graphical model. The experimental section demonstrates the effectiveness of the approach and the benefit of combining different contextual sources by evaluating the proposed system on the public database IAM-OnDo.

2. Related works

Over the last decade, much effort has been devoted to the interpretation of online hand-drawn content, from handwritten text [1,3], symbols and gestures [4–6], to structured scientific notations [7], diagrams and sketches [8–10] or tables[11,12]. The problem of analysing unconstrained, free-form online documents has also been receiving an increasing attention lately [13,14], and methods have been proposed for processing content block segmentation [15], text lines segmentation [16] or for document structure retrieval [13,17,18].

An online note-taking document is a sequence of points acquired along the trajectory of the pen on the surface, where points are organised in strokes (portions of the trajectory realised without lifting the pen from the surface). Strokes offer a natural unit for partitioning a document, and it is generally assumed that a stroke can be affected a single type (either text or non-text), as users usually lift the pen from the surface when switching from text to non-text content [14]. In the literature, the task of separating textual from non-textual strokes is regarded as a central problem for document understanding, whether it is considered in isolation (as in [2,13,14,19,20]) or combined with the inter-related tasks of segmentation and structure analysis (as in [15,17]).

Several categories of approaches for text/non-text stroke classification can be distinguished according to how contextual information is exploited. In this section, we first present approaches for isolated stroke classification, without any contextual information or simply including a local context description. We then present structured prediction methods, where interactions between elements are exploited to support the stroke classification task. We distinguish methods relying on the temporal structure of the document and methods relying on the spatial structure.

2.1. Isolated stroke classification

The work of Jain et al. [13] introduces a strictly local approach for classifying online strokes as text or non-text in handwritten documents. Two features are extracted from each stroke (namely the stroke curvature and stroke length), without considering interactions with other strokes. A linear classifier predicts the type of each stroke in isolation with a high accuracy (97%) and homogeneous regions such as text blocks or tables are detected in a subsequent step. The same features applied with a support vector machine classifier for isolated stroke classification on the challenging IAM-OnDo database were reported to perform significantly lower, at 91.3% [21]. Other authors have designed entropy-based methods, assuming that higher entropy of the online pen trajectory distinguishes the writing of text elements from the drawing of polygons or geometrical shapes [22,23]. Spectral features (e.g. discrete Fourier transform coefficients) have also been employed with a linear classifier for isolated stroke classification [24]. Willems et al. [25,26] propose to use a set of eight features (including global and structural features) with a nearest-neighbour classifier for detecting *modes* in online signal. In that case, the non-textual elements are further classified into subcategories considered as different modes (deictic gestures, complex drawings, geometric shapes).

If it makes sense to classify strokes individually, this decision problem can usually not to be answered unambiguously without considering some contextual information. For example, the same stroke shaped as a small circle could be considered to be a textual stroke (representing the letter “o”) or a non-text stroke (representing a circle) depending if it is located within a line of text or if it is a part of a drawing. Peterson et al. [27] thus proposed to extract features not only from the stroke to be classified but also from surrounding strokes. It was shown that extracting local context features (such as the average size of neighbouring strokes, or the average distance to them) significantly improves the stroke prediction. The findings of several other researchers [2,14] confirm this observation.

This clearly highlights the importance of considering contextual knowledge when predicting a stroke label. Accordingly, beyond inclusion of local context features from surrounding strokes, researchers have formulated the problem of stroke classification as a structured prediction problem, i.e. by modelling jointly the labelling of strokes from a document.

2.2. Exploitation of temporal context

An obvious way of modelling interactions between strokes in an online document is to exploit the temporal information. The online nature of the data suggests handling stroke labelling as a sequence labelling problem, where each stroke is an observation to be labelled as text or non-text.

As a direct exploitation of the temporal structure, the document can be segmented into sub-sequences of strokes that are assumed to share the same type because their temporal distance is below an empirical threshold. Then the classification can be operated at the subsequence level, by exploiting a richer context [28]. The problem of defining an appropriate segmentation strategy is not always clearly addressed [29] and the assumption of consistency of labels over sub-sequences is often unverified in realistic datasets [21].

A probabilistic framework for sequence prediction was adopted by Bishop et al. [20], who designed a hidden Markov model for modelling interactions between successive strokes in the drawing sequence. The dependencies express the fact that two strokes written successively are more likely to be of the same type. Emission probabilities for each stroke are computed by a multi layer perceptron with 11 input features and the transition probabilities are estimated from training data. Labelling strokes with the HMM model significantly outperforms the independent labelling with MLP classifier. The structured model can be improved with a bipartite HMM formulation, where the gaps between strokes are considered as observations and additional hidden states are integrated for modelling transitions between text and non-text states. A shortcoming of HMMs is the assumption of independence between observations [30], which in practice prevents from considering local context for prediction of a stroke label. In other words, the advantage of modelling dependencies between labels of temporally adjacent strokes is mitigated by the limitation to a weaker description for each stroke label prediction.

More recently, Indermuhle et al. [31] presented a mode detection approach based on bidirectional long short-term memory (BLSTM) neural networks. BLSTM is a type of recurrent neural networks that have been applied successfully to sequence prediction in speech and handwriting recognition [32]. For text detection in online documents, BLSTM is applied on the document represented as a stream of feature vectors extracted from points of the sampled pen trajectory. Prediction of labels is influenced by