Pattern Recognition 48 (2015) 1211–1224

Contents lists available at ScienceDirect

Pattern Recognition

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

# Tensor representation learning based image patch analysis for text identification and recognition

Guoqiang Zhong [a,*], Mohamed Cheriet [b]

[a] Department of Computer Science and Technology, Ocean University of China, Qingdao 266100, China
[b] Synchromedia Laboratory for Multimedia Communication in Telepresence, École de Technologie Supérieure, Montréal H3C 1K3, Canada

ARTICLE INFO

*Article history:*
Received 1 August 2013
Received in revised form
23 September 2014
Accepted 27 September 2014
Available online 24 October 2014

*Keywords:*
Tensor representation learning
Convergence
Ancient document understanding
Text identification
Text recognition

ABSTRACT

In this paper, we introduce a novel framework for text identification and recognition, called tensor representation learning based image patch analysis (TRL-IPA). Unlike most of previous text identification approaches, which can only be applied to binarized images, TRL-IPA can be directly applied to *gray level* and *color* images. TRL-IPA is built on a general formulation of the *convergent tensor representation learning* (CTRL) algorithms. In the implementation of TRL-IPA, image patches are represented in the form of tensors, while low dimensional representations of these tensors are learned via a CTRL algorithm. To identify text regions in new coming document images, a *random forest* classifier is trained in the learned tensor subspace. Moreover, the TRL-IPA framework can be straightforwardly applied to recognition problems, such as handwritten digits recognition. We conducted extensive experiments on *ancient Chinese, Arabic and Cyrillic* document images, to evaluate TRL-IPA on text identification tasks. Experimental results demonstrate its effectiveness and robustness. In addition, recognition results on images of handwritten digits show its advantage over state-of-the-art *vector and tensor representation* based approaches.



## 1. Introduction

Text identification is one of the most fundamental issues in several related areas, including layout analysis, text line segmentation, and document image understanding. Accurately detected text blocks are crucial and beneficial to the subsequent document analysis procedures, such as recognition and retrieval. Over the past few decades, many approaches for text identification have been proposed [1–13]. From the perspective of algorithmic strategy, most of these approaches can be classified into two categories: top-down and bottom-up approaches. In general, top-down approaches start from the whole image, and then segment it recursively into sub-regions [9]. On the contrary, bottom-up approaches start from combining pixels together, and then detect words, text lines and paragraphs subsequently [11]. These approaches have been successfully applied to printed document images, such as scanned newspapers, reports and journals. However, few methods are tested on ancient handwritten document images. Moreover, most of existing approaches can only be applied to binarized document images. It is difficult to extend them to applications involving gray level and color images.

In this work, we focus on the analysis of ancient handwritten document images. Because ancient handwritten document images convey lots of historical and cultural information from the past, they are valuable heritage. However, due to their irregularity, i.e. different scripts, different writing styles, additional notes on the text, fluctuating text lines, and the state of low quality, automatic processing ancient document images is a fairly challenging task. We show three example images in Fig. 1, to present some challenges of text identification on ancient handwritten document images. Fig. 1(a) shows an ancient Chinese document image.[1] We can see that the orientation of the text lines is vertical, and the sizes of the text blocks are quite different. Moreover, watermark symbols may arise further challenge to the text identification task. Fig. 1(b) shows an ancient Arabic document image,[2] which includes some paragraphs of notes besides the main text blocks. As a result, in this image, we can see many text blocks with different sizes and line orientations. Fig. 1 (c) shows an ancient Cyrillic document image.[3] Except for text blocks, some pictures and decorative letters were drawn in this image. To automatically distinguish between pictures and texts with high accuracy, it's a very challenging task for learning machines.

* Corresponding author.
*E-mail addresses:* gqzhong@ouc.edu.cn (G. Zhong), mohamed.cheriet@etsmtl.ca (M. Cheriet).

http://dx.doi.org/10.1016/j.patcog.2014.09.025
0031-3203/

[1] http://www.lib.ntnu.edu.tw/index.jsp
[2] http://wamcp.bibalex.org
[3] http://eng.digital.nb.rs


**Fig. 1.** Example images of ancient handwritten Chinese, Arabic and Cyrillic documents (zooming in for better vision of the scripts' details). (a) Chinese document image. (b) Arabic document image. (c) Cyrillic document image.

In this paper, we propose a novel framework for text identification, called tensor representation learning based image patch analysis (TRL-IPA). Unlike most of the previous text identification approaches, which can only be applied to binarized document images, TRL-IPA can be directly used on gray level and color document images. In contrast to the strategies of either top-down or bottom-up, TRL-IPA is based on tensor representation learning of the image patches. Particularly, it is built on a general formulation of the convergent tensor representation learning (CTRL) algorithms. Image patches are represented as tensors, and compact representations of these tensors are learned via a CTRL algorithm. To identify text regions in new coming document images, a random forest classifier is trained in the learned tensor subspace. To the end, the adjacent identified text patches are *automatically* combined to reveal the text regions in the document image. Moreover, small text regions may be discovered within only one image patch or several adjacent ones. In the following, we highlight some advantages of TRL-IPA for the text identification tasks.

- It can be applied to many kinds of document images, either printed or handwritten, either modern or ancient.
- It can be applied to gray level and color image. No image binarization is necessary, which may introduce noise and various artifacts.
- It does not assume any a priori knowledge about the layout of the document.
- It is insensitive to the orientation of the text lines.
- It is robust with respect to the algorithmic parameters.

The TRL-IPA framework can be straightforwardly applied to recognition problems, as long as the input images are represented as tensors. Subsequently, low dimensional representations of the tensors can be learned via a CTRL algorithm, while a random forest classifier can be employed to recognize new coming data.

The rest of this paper is organized as follows. We review some related work on text identification and recognition, as well as tensor representation learning in Section 2. Section 3 presents a general formulation of the CTRL algorithms. Its relationship with existing approaches and optimization algorithm are also presented in detail. In Section 4, we introduce the TRL-IPA framework and its appealing features for text identification and recognition. Section 5 shows the experimental results on text identification and recognition problems, where we compared TRL-IPA with related approaches. In Section 6, we conclude this paper with remarks and future work.

## 2. Related work

In this section, we review the existing approaches on text identification, text recognition, and tensor representation learning. The advantage and disadvantage of these approaches are analyzed, which motivate us to exploit convergent tensor representation learning models for text identification and recognition on gray level and color document images.

### 2.1. Text identification approaches

Text identification is a non-trivial task. Numerous approaches to text identification have been proposed in the literature. As mentioned in Section 1, nearly all of them can be basically classified into two categories: top-down and bottom-up approaches. In the following, we review some classic methods, among others.

One of the widely used top-down techniques is based on the run length smoothing (RLS) algorithm [1] and projection profiles [14]. The document image is smoothed via RLS, and projection profiles are utilized to identify the text blocks. After detection of the text blocks, one may further split the blocks into columns, paragraphs, text lines and words by using similar algorithm [14]. Based on the white gap between the blocks, Krishnamoorthy et al. [15] proposed a recursive $X$–$Y$ cut algorithm, which stores the identification process in a labeled $X$–$Y$ tree. Ittner and Baird [16] proposed a robust and multi-lingual top-down method, which detects the text blocks after the skew correction of the images. This method generates the minimal spanning tree for finding the text line orientation, and utilizes the projection profiles of the blocks to identify the text lines. In general, top-down approaches are fast, since finding the white gaps is a linear algorithm. However, most of these methods work well only for document images with Manhattan layout, i.e. with clear horizontal and vertical white gaps between the segments. To the end, for ancient document images with complicated layout, such as those shown in Fig. 1(b) and (c), top-down approaches may fail to identify the text blocks correctly.

In contrast to the above, bottom-up approaches use local information to build up higher level units. O'Gorman [17] proposed the *docstrum* method, which merges neighboring connected components using rules based on the geometric relationship between $k$ nearest neighbor pairs. The author shows that *docstrum* can perform well on printed document images as well as handwritten document images with slightly curved lines. Tsujimoto and Asada [18] proposed a document analysis method, based on the run-length image representation. Words are firstly extracted from a document image, and then merged into text lines. Subsequently, text lines are combined into blocks. Based on the run-length smearing algorithm (RLSA) [19], Fan et al. [20] proposed a document analysis system. RLSA has the effect of linking together neighboring black areas that are separated by less than $L$ pixels, where $L$ is a predefined number. After RLSA being performed, the close text lines are merged and the resulting blocks are classified according to a feature based classification scheme. Generally speaking, bottom-up methods are widely applicable to various