Contents lists available at ScienceDirect

Pattern Recognition

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

Improving patch-based scene text script identification with ensembles of conjoined networks

Lluis Gomez*, Anguelos Nicolaou, Dimosthenis Karatzas

Computer Vision Center, Universitat Autonoma de Barcelona. Edifici O, Campus UAB, 08193 Bellaterra (Cerdanyola) Barcelona, Spain

ARTICLE INFO

Article history:
Received 12 May 2016
Revised 1 December 2016
Accepted 30 January 2017
Available online 3 February 2017

Keywords:
Script identification
Scene text understanding
Multi-language OCR
Convolutional neural networks
Ensemble of conjoined networks

ABSTRACT

This paper focuses on the problem of script identification in scene text images. Facing this problem with state of the art CNN classifiers is not straightforward, as they fail to address a key characteristic of scene text instances: their extremely variable aspect ratio. Instead of resizing input images to a fixed aspect ratio as in the typical use of holistic CNN classifiers, we propose here a patch-based classification framework in order to preserve discriminative parts of the image that are characteristic of its class.

We describe a novel method based on the use of ensembles of conjoined networks to jointly learn discriminative stroke-parts representations and their relative importance in a patch-based classification scheme. Our experiments with this learning procedure demonstrate state-of-the-art results in two public script identification datasets.

In addition, we propose a new public benchmark dataset for the evaluation of multi-lingual scene text end-to-end reading systems. Experiments done in this dataset demonstrate the key role of script identification in a complete end-to-end system that combines our script identification method with a previously published text detector and an off-the-shelf OCR engine.

1. Introduction

Script and language identification are important steps in modern OCR systems designed for multi-language environments. Since text recognition algorithms are language-dependent, detecting the script and language at hand allows selecting the correct language model to employ [1]. While script identification has been widely studied in document analysis [2,3], it remains an almost unexplored problem for scene text. In contrast to document images, scene text presents a set of specific challenges, stemming from the high variability in terms of perspective distortion, physical appearance, variable illumination and typeface design. At the same time, scene text comprises typically a few words, contrary to longer text passages available in document images.

Current end-to-end systems for scene text reading [4–6] assume single script and language inputs given beforehand, i.e. provided by the user, or inferred from available meta-data. The unconstrained text understanding problem for large collections of images from unknown sources has not been considered up to very recently [7–11]. While there exists some previous research in script identification of text over complex backgrounds [12,13], such methods have been so far limited to video overlaid-text, which presents in general different challenges than scene text.

This paper addresses the problem of script identification in natural scene images, paving the road towards true multi-lingual end-to-end scene text understanding Fig. 1. Multi-script text exhibits high intra-class variability (words written in the same script vary a lot) and high inter-class similarity (certain scripts resemble each other). Examining text samples from different scripts, it is clear that some stroke-parts are quite discriminative, whereas others can be trivially ignored as they occur in multiple scripts. The ability to distinguish these relevant stroke-parts can be leveraged for recognizing the corresponding script. Fig. 2 shows an example of this idea.

The use of state of the art CNN classifiers for script identification is not straightforward, as they fail to address a key characteristic of scene text instances: their extremely variable aspect ratio. As can be seen in Fig. 3, scene text images may span from single characters to long text sentences, and thus resizing images to a fixed aspect ratio, as in the typical use of holistic CNN classifiers, will deteriorate discriminative parts of the image that are characteristic of its class. The key intuition behind the proposed method is that in order to retain the discriminative power of stroke parts we must rely in powerful local feature representations and use them within a patch-based classifier. In other words, while holistic CNNs have superseded patch-based methods for image classification, we claim

* Corresponding author.
E-mail addresses: lgomez@cvc.uab.cat (L. Gomez), anguelos@cvc.uab.es (A. Nicolaou), dimos@cvc.uab.cat (D. Karatzas).

http://dx.doi.org/10.1016/j.patcog.2017.01.032
0031-3203/© 2017 Elsevier Ltd. All rights reserved.

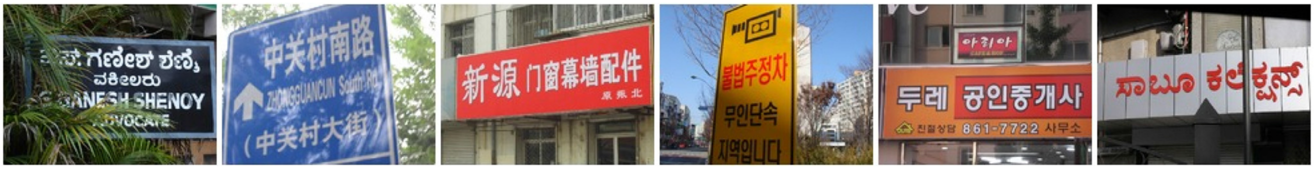

Fig. 1. Collections of images from unknown sources may contain textual information in different scripts.

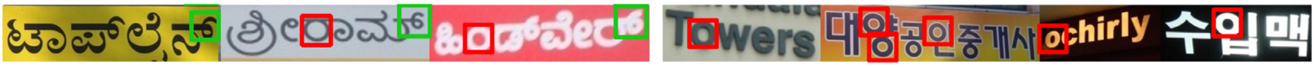

Fig. 2. (best viewed in color) Certain stroke-parts (in green) are discriminative for the identification of a particular script (left), while others (in red) can be trivially ignored because are frequent in other classes (right). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

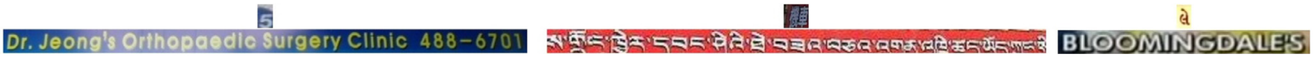

Fig. 3. Scene text images with the larger/smaller aspect ratio available in three different datasets: MLe2e(left), SIW-13(center), and CVSI(right).

that patch-based classifiers can still be essential in tasks where image shrinkage is not feasible.

In previously published work [10] we have presented a method combining convolutional features, extracted by sliding a window with a single layer Convolutional Neural Network (CNN) [14], and the Naive–Bayes Nearest Neighbor (NBNN) classifier [15] with promising results. In this paper we demonstrate far superior performance by extending our previous work in two different ways: First, we use deep CNN architectures in order to learn more discriminative representations for the individual image patches; Second, we propose a novel learning methodology to jointly learn the patch representations and their importance (contribution) in a global image to class probabilistic measure. For this, we train our CNN using an Ensemble of Conjoined Networks and a loss function that takes into account the global classification error for a group of *N* patches instead of looking only into a single image patch. Thus, at training time our network is presented with a group of *N* patches sharing the same class label and produces a single probability distribution over the classes for all them. This way we model the goal for which the network is trained, not only to learn good local patch representations, but also to learn their relative importance in the global image classification task.

Experiments performed over two public datasets for scene text classification demonstrate state-of-the-art results. In particular we are able to reduce classification error by 5 percentage points in the SIW-13 dataset. We also introduce a new benchmark dataset, namely the MLe2e dataset, for the evaluation of scene text end-to-end reading systems and all intermediate stages such as text detection, script identification and text recognition. The dataset contains a total of 711 scene images, and 1821 text line instances, covering four different scripts (Latin, Chinese, Kannada, and Hangul) and a large variability of scene text samples.

2. Related work

Script identification is a well-studied problem in document image analysis. Gosh et al. [2] has published a comprehensive review of methods dealing with this problem. They identify two broad categories of methods: structure-based and visual appearance-based techniques. In the first category, Spitz and Ozaki [16,17] propose the use of the vertical distribution of upward concavities in connected components and their optical density for page-wise script identification. Lee et al. [18], and Waked et al. [19] among others build on top of Spitz seminal work by incorporating additional connected component based features. Similarly, Chaudhuri et al. [20] use the projection profile, statistical and topological features, and stroke features for classification of text lines in printed documents. Hochberg et al. [21] propose the use of cluster-based templates to identify unique characteristic shapes. A method that is similar in spirit with the one presented in this paper, while requiring textual symbols to be precisely segmented to generate the templates.

Regarding segmentation-free methods based on visual appearance of scripts, i.e. not directly analyzing the character patterns in the document, Wood et al. [22] experimented with the use of vertical and horizontal projection profiles of full-page document images. More recent methods in this category have used texture features from Gabor filters analysis [23–25] or Local Binary Patterns [26]. Neural networks have been also used for segmentation-free script identification [27,28] without the use of hand-crafted features.

All the methods discussed above are designed specifically with printed document images in mind. Structure-based methods require text connected components to be precisely segmented from the image, while visual appearance-based techniques are known to work better in bilevel text. Moreover, some of these methods require large blocks of text in order to obtain sufficient information and thus are not well suited for scene text which typically comprises a few words.

Contrary to the case of printed document images, research in script identification on non traditional paper layouts is more scarce, and has been mainly dedicated to handwritten text [29–33], and video overlaid-text [12,13,34–36] until very recently. Gllavatta et al. [12], in the first work dealing with video text script identification, proposed a method using the wavelet transform to detect edges in overlaid-text images. Then, they extract a set of low-level edge features, and make use of a K-NN classifier.

Sharma et al. [34] have explored the use of traditional document analysis techniques for video overlaid-text script identification at word level. They analyze three sets of features: Zernike moments, Gabor filters, and a set of hand-crafted gradient features previously used for handwritten character recognition. They propose a number of pre-processing algorithms to overcome the inherent challenges of video overlaid-text. In their experiments the combination of super resolution, gradient features, and a SVM classifier perform significantly better that the other combinations.

Phan et al. [35] propose a method for combined detection of video text overlay and script identification. They propose the extraction of upper and lower extreme points for each connected component of Canny edges of text lines and analyze their smoothness and cursiveness.