Pattern Recognition Letters 100 (2017) 6–13

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Robust shared feature learning for script and handwritten/machine-printed identification

Ziyong Feng[a,1], Zhaoyang Yang[a,1], Lianwen Jin[a], Shuangping Huang[a,*], Jun Sun[b]

[a] School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China
[b] Fujitsu R&D Center Co., Ltd., Beijing, China

ARTICLE INFO

Article history:
Received 25 August 2016
Available online 8 September 2017

ABSTRACT

In this paper, we focus on the problem of script and handwritten/machine-printed identification of texts. We simultaneously identify the script (Chinese, English, Japanese, Korean, or Russian) and whether it is handwritten or machine-printed text by designing a dual-branch structured deep convolutional neural network (CNN). For the training stage, we propose a two-stage multi-task learning strategy to learn robust shared features for script and handwritten/machine-printed identification. Accordingly, we can implement two identification tasks using the proposed single CNN model. We compare the effects of using different length of input to train CNN. The experimental results show that text-line input is a suitable choice for the two identification tasks, as it can effectively capture more discriminative content for both script and handwritten/machine-printed identification. Furthermore, we evaluate three CNN networks of different scales (small, medium, and large) to determine the best CNN architecture for script and handwritten/machine-printed identification. As shown by our experimental validation, integrating the text-line input with larger architecture significantly improves performance. The accuracies achieved by the two-stage multi-task CNN for handwritten/machine-printed and script identification are 99% and 95%, respectively.

1. Introduction

With the rapidly growing number of digitized documents, automatic document analysis is becoming increasingly important. Machine-printed and handwritten texts in different languages can coexist on the same page in many practical documents, such as bank checks, forms, and letters [1,2]. Discriminating between machine-printed and handwritten texts and classifying text language are necessary for an automatic document analysis system because different types of texts require different recognition engines. Most optical character reader (OCR) systems recognize machine-printed and handwritten text using different methods. In addition, there are different recognition systems for different language texts. For example, an English document is typically recognized on a text-line or word level by a sequential classifier, e.g., hidden Markov model/long short-term memory (HMM/LSTM) [3]. However, a Chinese document must be segmented into characters and sub-characters. These characters are recognized separately, and a language model is employed to generate the final result for the entire document [4]. Moreover, category information facilitates document retrieval.

Three categories of text identification have been considered in previous studies, specifically, text-line, word, and character levels. For text-line-level identification, Jlaiel et al. [5] presented a three-decision-level strategy for Arabic and Latin text identification of machine-printed and handwritten natures. In addition, Kavallieratou and Stamatatos [6] designed a feature extraction method using a horizontal histogram of upper and lower profiles as well as structural information. A quadratic classifier was employed to discriminate between machine-printed and handwritten texts. Furthermore, Pal and Chaudhuri [7] proposed structural and statistical features to distinguish machine-printed text lines from handwritten text lines in Bangla and Devnagari. Nicolaou et al. [8] proposed a script identification method based on handcrafted texture features and an artificial neural network for video text and handwritten text.

For word-level identification, Zheng et al. [1] converted the two-class identification problem into a three-class problem by adding auxiliary class noise. Furthermore, Mozaffari and Bahar [9] applied a structural and cross-counting histogram feature with a k-nearest neighbor/support vector machine (KNN/SVM) classi-

* Corresponding author.
E-mail address: eehsp@scut.edu.cn (S. Huang).
[1] Ziyong Feng and Zhaoyang Yang contributed equally to this work.

http://dx.doi.org/10.1016/j.patrec.2017.09.016
0167-8655/

Table 1
Architectures of three SH–CNNs.

Layer Type	Small	Medium	Large
Convolution	32 × 5 × 5 stride 1, pad 2	32 × 5 × 5 stride 1, pad 2	96 × 4 × 4 stride 1, pad 0
Pooling	2 × 2 maxpool	2 × 2 maxpool	2 × 2 maxpool
Convolution	64 × 5 × 5 stride 1, pad 2	64 × 5 × 5 stride 1, pad 2	128 × 4 × 4 stride 2, pad 2
Pooling	2 × 2 maxpool	2 × 2 maxpool	2 × 2 maxpool
Convolution		64 × 5 × 5 stride 1, pad 2	256 × 2 × 2 stride 1, pad 1
Convolution			256 × 2 × 2 stride 1, pad 0
Convolution			256 × 2 × 2 stride 1, pad 1
Pooling		2 × 2 maxpool	2 × 2 maxpool
Fully connected	512 256	512 256	1024 512
Fully connected	256	256	512
Softmax	5 2	5 2	5 2

fier for Farsi/Arabic text identification. In addition, Saïdani et al. [10] focused on Arabic and Latin words.

For character-level identification, Kuhnke et al. [11] proposed a system for identifying machine-printed and handwritten characters using structural feature extraction and a neural network classifier. Fan et al. [12] detected text blocks and used character block histograms for Chinese and English texts.

Despite the above contributions, existing methods are limited to a single text category. The methods for word-level and character-level identification require precise segmentation. Moreover, most approaches concentrate on a single task: handwritten/machine-printed text discrimination or script identification. Only a few of them [5,7,10] handle both two tasks simultaneously. Nevertheless, these methods only examine datasets that contain no more than two languages. In fact, it is not easy to directly adapt these methods to multi-language cases. In addition, existing identification systems perform hand-crafted feature extraction and classification separately. Recently, numerous end-to-end approaches, such as the convolutional neural network (CNN) [13], have been able to jointly learn the features and classifier. Therefore, the CNN is an appropriate model for script and handwritten/machine-printed identification. Although CNN mostly outperforms the traditional methods, huge computation is required. While two separate CNNs need two forward passes, a dual-branch of two related tasks just requires one. For example, the FLOPs (multiply-adds) of the medium model in Table 1 (as referred in Section 3) is 85.20 M. If we apply two separate networks for the two tasks of script and handwritten/machine-printed identification, the FLOPs increases to 167.12 M. In contrast, a dual-branch CNN for these two tasks can reduce computation by sharing features between them, which makes the whole identification system more efficient.

In this paper, we identify the script and handwritten/machine-printed from a dataset containing five languages by designing a dual-branch CNN. The feasibility of the dual-branch structure enables the two-stage multi-task learning framework to learn the robust shared features. In addition, we verify the effect of inputs with different length for deep CNN. The experimental results show that text-line input is more suitable than square input, because the text-line input can effectively capture more discriminative content for both script and handwritten/machine-printed identification. Finally, we evaluate three types of CNN architectures for script and handwritten/machine-printed identification to analyze the relationship between the model size and the performance.

The remainder of this paper is organized as follows. Section 2 describes the comparison of inputs with different length for CNN. Section 3 introduces the two-stage multi-task learning framework for script and handwritten/machine-printed identification. Section 4 presents and analyzes the experimental results. Finally, Section 5 concludes the paper.

2. Text-line input

CNNs with fully connected layers require inputs with fixed size. However, text-line images have different length, which cannot be straightforward used as the inputs of the CNNs with full connection layer. Furthermore, resizing the whole dataset to a uniform size is not feasible as the aspect ratios of the text-line images vary greatly and are mostly much smaller than 1. If we crudely resize the text-line images, lots of information would be missing, which is detrimental to both script and handwritten/machine-printed identification.

2.1. Text-line input for CNN

Intuitively, we should segment all characters in the text image and warp these characters to a fixed size for input to the CNN. However, it is difficult to precisely segment characters because the segmentation heavily depends on low-level image processing operations, such as image binarization and edge detection. Due to multiple resolutions, background clutter, lighting, and noise, image binarization and edge detection may lead to conglutination of some characters. In such cases, several characters and strokes appear in a segmented region, especially for Latin text images. If we resize these regions in a straightforward manner, all characters and strokes are compressed along the vertical direction. Discriminative information is lost in a manner similar to that when resizing the entire text image.

Therefore, a patch-based CNN is a feasible system for overcoming this problem. Let the size of a text-line image be $h \times w$. Then, we can randomly crop a square patch of size $h \times h$. A CNN (*sq*-CNN) can be trained using these fixed-size square input patches. The formulation of convolutional layers is given by

$$\mathbf{x}_j^l = \sigma\left(\sum_i \mathbf{x}_i^{l-1} * \mathbf{k}_{ij}^l + b_j^l 1_{(M_{l-1}-K_l+1)\times(M_{l-1}-K_l+1)}\right), \tag{1}$$

where $\mathbf{x}_i^{l-1} \in \mathbb{R}^{M_{l-1}\times M_{l-1}}$ is the ith feature map in layer $l-1$, $\mathbf{k}_{ij}^l \in \mathbb{R}^{K_l\times K_l}$ is the filter connected with the ith feature map in layer $l-1$ and the jth feature map in layer l, and b_j^l is the bias of the jth feature map in layer l. $*$ denotes the convolution operation, and $\sigma(x) = \max(0, x)$ is an activation function that uses rectified linear units (ReLUs) [16] instead of the sigmoid function or hyperbolic tangent function. For this activation function with non-saturating nonlinearity, it has been shown that the error rate can be rapidly decreased in a training step using the stochastic gradient descent (SGD) [15].

However, regardless of whether we consider script or handwritten/machine-printed identification, the discriminative features in a text-line image are mostly hidden in the vertical direction. Thus, the square input patch obtained by resizing may suffer from loss of information. To address this problem, we