Pattern Recognition 63 (2017) 397–405

Contents lists available at ScienceDirect

Pattern Recognition

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

Accurate recognition of words in scenes without character segmentation using recurrent neural network

Bolan Su, Shijian Lu*

Institute for Infocomm Research (I2R), #21-01, 1 Fusionopolis Way, Singapore 138632, Singapore

ARTICLE INFO

Keywords:
Scene text recognition
Recurrent neural network

ABSTRACT

Recognition of texts in scenes is one of the most important tasks in many computer vision applications. Though different scene text recognition techniques have been developed, scene text recognition under a generic condition is still a very open and challenging research problem. One major factor that defers the advance in this research area is character touching, where many characters in scene images are heavily touched with each other and cannot be segmented for recognition. In this paper, we proposed a novel scene text recognition technique that performs word level recognition without character segmentation. Our proposed technique has three advantages. First it converts each word image into a sequential signal for the scene text recognition. Second, it adapts the recurrent neural network (RNN) with Long Short Term Memory (LSTM), the technique that has been widely used for handwriting recognition in recent years. Third, by integrating multiple RNNs, an accurate recognition system is developed which is capable of recognizing scene texts including those heavily touched ones without character segmentation. Extensive experiments have been conducted over a number of datasets including several ICDAR Robust Reading datasets and Google Street View dataset. Experiments show that the proposed technique is capable of recognizing texts in scenes accurately.

1. Introduction

Text recognition in scenes is one of the most important research areas in computer vision and it has been studied for many years with different successful applications. Due to the rapid development of mobile sensors and internet technology, a huge amount of digital images are produced every day. Textual regions as one of the most informative regions in scene images need to be interpreted properly and automatically to make these images more accessible and valuable.

The Robust Reading Competitions [1,2] held under the framework of the International Conference on Document Analysis and Recognition(ICDAR) 2011 & 2013 show recent development on this research topic. One of tasks in these competitions is to recognize cropped word images which have little constraints in terms of text fonts, environmental lighting, image background, etc. A number of recognition systems have been reported and evaluated over the benchmarking datasets and the recognition accuracy has been lifted from the initial around 50% to the recent around 80% over the last decades.

Scene text recognition has been investigated in two typical approaches. The first is the traditional OCR (Optical Character Recognition) approach, which first segments text pixels from the image background and then applies some existing OCR engine to recognize the segmented characters. Another is feature based approach, which extracts various visual features such as HOG (histograms of oriented gradients) and SIFT (scale-invariant feature transform) to train a multi-class character classification model.

The traditional OCR techniques have been developed for decades and achieved great success in different commercial systems. On the other hand, most of them are designed for the scanned document texts which are usually well formatted and have a good image quality. They often fail to produce good results when applied for texts in scenes, where characters have little constraints in term of text fonts, environmental lighting, image background, etc. as illustrated in Figs. 1 (a) and (e). Several systems [3–5] have been reported to extract a clean character regions before feeding to OCR engines but they usually suffer from two typical constraints. First, text segmentation in scene images is a non-trivial problem due to uneven illumination, blur, low text background contrast, etc., as illustrated in Figs. 1 (e), and (g). Second, texts in scene images often have perspective distortion and special fonts, which cannot be recognized by traditional OCR engines properly as illustration in Figs. 1 (c) and (h). Different image restoration techniques [6,7] are often required to produce satisfactory recognition results.

* Corresponding author.
E-mail addresses: subl@i2r.a-star.edu.sg (B. Su), slu@i2r.a-star.edu.sg (S. Lu).

http://dx.doi.org/10.1016/j.patcog.2016.10.016
Received 6 June 2016; Received in revised form 20 September 2016; Accepted 15 October 2016
Available online 22 October 2016
0031-3203/

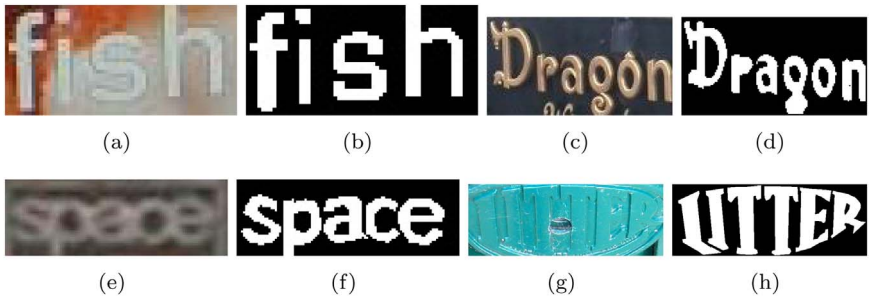

Fig. 1. Four text image examples and their corresponding text segmentation ground truth that are taken from the benchmarking word image dataset [8] (From up to down, the text images become more difficult to recognize). The OCR results obtained using Abbyy Fine Reader 10.0 are (a: r), (b: fish), (c: –), (d: Draoon), (e: –), (f: –), (g: –), (h: –), where '–' denotes no results produced.

Fig. 2. Word image examples taken from the recent Public Datasets [22,1,2]. All the words in the images are correctly recognized by our proposed method.

The other approach exploits the object recognition techniques that have been extensively studied in recent years. In particular, these techniques can be categorised into two groups, namely character level recognition methods [9–16] and word level recognition methods [17–19]. The character level recognition methods first recognize each character of the word image, and then group all the recognized characters into a word string. Various visual features such as HOG [12,13,20], and part based tree structure [14] have been exploited to represent characters in scenes. The convolutional neural network (CNN) has also been widely used as the character classifier in recent years [9–11,16]. Besides, different clustering strategies have been proposed to group the recognized characters into a word string such as pictorial structure [13], conditional random field [12,14], HMM [9], N-gram model [10,16], etc. On the other hand, segmenting a word images into character images is often a very challenging task and sometime even impossible as illustrated in Fig. 2 [21].

The word-level recognition treats each word image as a whole and performs recognition without the character segmentation. Different techniques [17–19] have been proposed in recent years and very promising results have been obtained. In particular, the discrete wavelet transform (DWT) method [17] tries to find smallest distance between the word images and the font-renderings words within a lexicon. The attribute embedding method [18] creates a joint embedding space for word images and the word strings within a lexicon and finds a close match. The Whole Word Deep CNN method [19] treats each possible word in the lexicon as an output label of the trained CNN. The common limitation of these methods is that they all require an explicit lexicon which is costly and often inaccessible under many scenarios.

In [23], we proposed a scene text recognition technique that treats a word image as an unsegmented sequence. The major advantage is that it does not require an explicit lexicon (e.g. all the possible words are listed) and can perform the word-level recognition without lexicon or with an implicit lexicon (e.g. some constraints on the output word string) which is much easier to construct. Input images are normalized into the same height and retain the aspect ratio before the feature extraction. The column feature is extracted by using a fixed sized window. The major limitation of [23] is that the column features with a fixed window size cannot capture characteristics of different characters concurrently. The reason is that the aspect ratio of different characters such as 'i′, 'I′, 'W′, and 'M′ is very different, and so the same character in different fonts.

The new model as presented in this paper addresses the limitation and improves the word recognition accuracy significantly. In particular, we used image patches of different sizes to handle the large character aspect ratio variation and this approach also captures much richer characteristics of texts. Generally speaking, a small image patch can capture the stroke-level features as well as those thin characters such as 'l′ and 'i′, whereas a larger image patch is able to capture the character/intra-character level features as well as those wide characters such as 'M′ and 'W′. In addition, the new model implemented multiple recurrent neural networks (RNNs) to combine column features from patches of different window sizes. Experiments show that the new model is robust and able to recognize various challenging word images correctly.

The contributions can be summarized as follows:

- First, we design an effective way of converting a word image into a sequential signal so that RNN techniques, which have been successfully used in speech processing and handwriting recognition areas, can be introduced and applied. We adapt RNN for the recognition of texts in scenes, and design a segmentation-free scene word recognition system that obtains superior word recognition accuracy.
- Second, we propose a new ensembling technique that combines outputs from two RNNs for better recognition results. The proposed ensembling technique is generic and can be easily extent to ensemble other models for better performance.
- Third, compared with some systems [10,11] that rely heavily on certain local dataset (which are not available to the public), our system makes use of several publicly available datasets in training