Neurocomputing 184 (2016) 78–90

Contents lists available at ScienceDirect

Neurocomputing

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

Convolutional feature learning and Hybrid CNN-HMM for scene number recognition

Qiang Guo *, Fenglei Wang, Jun Lei, Dan Tu, Guohui Li

College of Information System and Management, National University of Defense Technology, Changsha 410072, China

ARTICLE INFO

Article history:
Received 22 January 2015
Received in revised form
20 May 2015
Accepted 13 July 2015
Available online 17 December 2015

Keywords:
Scene text recognition
Convolutional neural network
Hidden Markov model
Deep learning
Hybrid NN-HMM
GMM-HMM

ABSTRACT

In this work, we investigate to recognize house numbers captured in street view images. We formulate the problem as sequence recognition and present an integrated model by combining Convolutional Neural Network (CNN) and Hidden Markov Model (HMM). Our method utilizes representation capability of CNN to model the highly variable appearance of digits. Meanwhile, HMM is used to handle the dynamics of the image sequence. They are combined in a hybrid way to form the Hybrid CNN-HMM. Using this model, we can perform training and recognition both at the whole image level without explicit segmentation. The model makes CNN applicable to dynamic problems. Experiments show that the Hybrid CNN-HMM can dramatically boost the performance of Gaussian Mixture Model (GMM)-HMM. We evaluate different local features, *e.g.* LBP, SIFT and HOG, as observations fed into HMM and find CNN features consistently surpass those hand-engineered features with respect to recognition accuracy. To gain insight into performance difference of the features, we map them from the high-dimensional space to a 2-D plane by the t-SNE algorithm to visualize their semantic clustering with respect to the task. The visualization clearly justified the efficiency of features learnt by CNN.

1. Introduction

Though the research of recognizing handwritten and machine printed characters lasts for several decades [1,2], recognizing text in natural scene still remains a difficult problem. Text captured in unconstrained natural scene shows a great deal of variability of appearance, *e.g.* it has different fonts, scales, rotations, and illumination conditions. In this paper, we address the problem of recognizing house numbers captured in street view images. We focus on image modeling rather than language modeling.

Traditional methods for text recognition usually fall into a two-stage pipeline of first segmenting the image to extract isolated characters then perform recognition on the extracted characters [3–5].

Segmentation and recognition have long been considered to be dependent on each other, which is known as Sayre's paradox [6,7]. The paradox makes it difficult to do either. Meanwhile, since numerous factors affect the character appearance, the segmentation stage is prone to failure when facing natural scene images [8]. Furthermore, we need a large amount of isolated character samples to cover the variability of their appearance. The work of exactly labeling every character is highly labor consuming.

Our aim is to unify segmentation and recognition. To this end, we formulate the problem as sequence recognition and propose the Hybrid CNN-HMM. The arrangement of digits in street view images is usually horizontal, occasionally oblique. Based on this observation, our model performs sliding window on the input image to extract a sequence of frames. In a generative view, we treat the frames as being produced by the transition of an underlying state sequence. In this situation, the states represent digits from 0 to 9 and background clutter. For our model, training and recognition are directly performed on the whole image level. Fig. 1 shows an overview of our model's architecture.

Starting with the milestone of ImageNet [9] classification results reported by Krizhevsky et al. [10] , deep CNN has shown its impressive performance on various vision tasks [11–14]. The high efficiency of CNN results from the strong capability for hierarchical feature learning given a large amount of data. Hand-engineered feature descriptors, such as SIFT [15], HOG [16] and LBP [17], afford an intuitive interpretation as histograms of oriented edge filter responses arranged in spatial blocks. Though those features are widely used in vision problems [18–20], they are not generated from an optimization process to be compatible with the specific problem, and insufficient to be encoded with supervision. It seems

* Corresponding author.
E-mail addresses: guoqiang05@nudt.edu.cn (Q. Guo), wangfenglei@nudt.edu.cn (F. Wang), leijun1987@gmail.com (J. Lei), tudan@nudt.edu.cn (D. Tu), guohli@nudt.edu.cn (G. Li).

http://dx.doi.org/10.1016/j.neucom.2015.07.135
0925-2312/

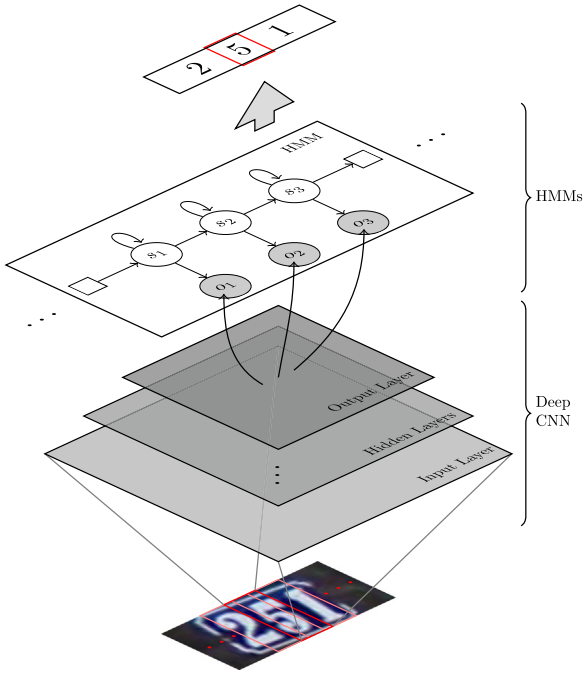

Fig. 1. Hybrid CNN-HMM. This image demonstrates the recognition process of our model. Input image is processed by sliding window to extract a sequence of frames. Each frame is normalized to the same scale and fed into the CNN which produce its posterior probability belonging to a category. This probability, after normalization, is used as the output probability of HMM. HMM is used to infer the most probable digits out of the frame sequence.

that CNN is replacing hand-engineered features for a wide variety of problems.

However, CNN is typically used as a static model whose input feature is fixed-dimensional. To apply CNN to sequence problem, we need to facilitate CNN with a dynamic temporal model. We choose HMM, which is a widely used model for sequence modeling. HMM-based methods have been the de facto standard for handwriting recognition [7,21] and speech recognition [22,23]. The two research areas are closely related to our problem.

Our idea is mostly motivated by the complementary modeling capacities of CNN and HMM. Specifically, HMM is used for temporal modeling and CNN deals with all kinds of character appearance variations. CNN is critical to the performance for its great representation capability. Meanwhile, the outputs of the softmax layer can be used as the *a posteriori* probabilities conditioned on certain hidden states [24].

Recently, Deep Neural Networks (DNNs) have been ubiquitously used in speech recognition [25–28] and achieved significant improvements compared with traditional GMM-HMM methods. Our model is similar with the CD-DNN-HMM [25] in speech recognition but uses CNN for vision problems.

To use CNN which is a fully supervised model, we have to give initial states assignment for each frame. We construct a GMM-HMM (Fig. 2) for bootstraping. The bootstrap process produces the initial frame-state assignments. We then switch to train the Hybrid CNN-HMM. Starting with a not-so-well performed GMM-HMM, we can gain dramatic performance improvements by replacing GMM with CNN as the observation model.

We evaluate different features fed into GMM-HMM and show that CNN features largely outperform several engineered local features. The features of CNN are directly learnt from word level images, and show good performance even with the dataset containing disturbed wrongly assigned samples. Experiments show

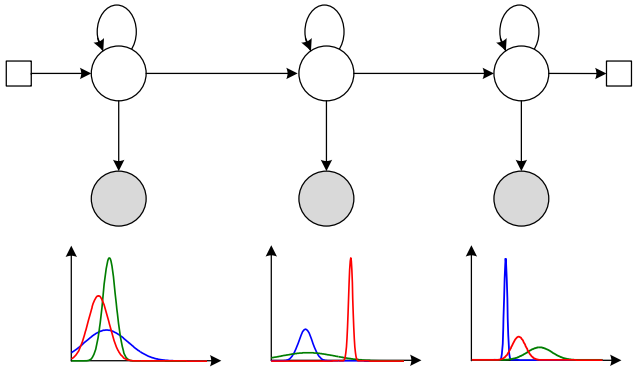

Fig. 2. GMM-HMM.

that CNN can dramatically boost the performance of GMM-HMM. We obtain competitive results on the street view house number (SVHN) dataset.

The rest of the paper is structured as follows. Section 2 briefly surveys related works in this field. Section 3 describes our formulation of the problem and an overview of the methods for training and recognition. Section 4 explains the details of the Hybrid CNN-HMM. Experimental results and comparisons are shown in Section 5. Conclusion remarks and potential directions for future research are presented in Section 6.

2. Related work

Scene number recognition can fall into the category of natural scene text recognition problem. Research on scene text recognition had started in mid-90s [29], but still remains a difficult problem. Different from machine-printed character or handwriting recognition problem, recognizing text in natural scene has its special difficulties. Text captured in unconstrained natural scene shows quite large appearance variability. They have different fonts, scales, rotations, lightening conditions and are organized in various kinds of layouts.

Traditional methods dealing with this problem are based on sequential character classification by either sliding window [30,31] or extracting connected components [32,33], after which a word prediction is made by grouping character classifier predictions in a left-to-right manner. More recent works [34–36] make use of over-segmentation guided by a supervised classifier to generate character candidates. Words are recognized through a sequential beam search optimization over character candidates. Usually, a small fixed lexicon is used as language model to constrain word recognition [30,31] in scene text recognition. However, the problem of recognizing street view numbers cannot benefit from language model, which makes the problem more difficult.

In all these systems, character model is the most critical component. Nowadays, CNN is showing more and more powerful capability in object recognition tasks [10,37–39]. Some works have used CNN to tackle scene text recognition problem [31,38,40,35]. Among these research, CNN shows great capability to represent highly variable character appearance in natural scene and still holding good discrimination capability.

An important problem is how to infer the characters from the whole image when it is difficult to isolate each single character from the image. Most of the previous work on SVHN dataset recognize cropped digits. Only a few works are done on the full numbers format of SVHN.

Convolutional networks have been successfully applied to recognition of handwritten numbers in 1990s [41,42]. Recently,