Scene Text Recognition using Higher Order Language Priors

Anand Mishra[1]
http://researchweb.iiit.ac.in/~anand.mishra/
Karteek Alahari[2]
http://www.di.ens.fr/~alahari/
C.V. Jawahar[1]
http://www.iiit.ac.in/~jawahar/

[1] CVIT
IIIT Hyderabad
Hyderabad, India
[2] INRIA - WILLOW
ENS
Paris, France

The problem of recognizing text in images taken in the wild has gained significant attention from the computer vision community in recent years. The scene text recognition task is more challenging compared to the traditional problem of recognizing text in printed documents. We focus on this problem, and recognize text extracted from natural scene images and the web. Significant attempts have been made to address this problem in the recent past, for example [1, 2]. However, many of these works benefit from the availability of strong context, which naturally limits their applicability. In this work, we present a framework to overcome these restrictions. Our model introduces a higher order prior computed from an English dictionary to recognize a word, which may or may not be a part of the dictionary. We present experimental analysis on standard as well as new benchmark datasets.

The main contributions of this work are: (1) We present a framework, which incorporates higher order statistical language models to recognize words in an unconstrained manner, *i.e.* we overcome the need for restricted word lists. (2) We achieve significant improvement (more than 20%) in word recognition accuracies in a general setting. (3) We introduce a large word recognition dataset (atleast 5 times larger than other public datasets) with character level annotation and benchmark it.

Method Overview. We propose a CRF based model for recognizing words. The CRF is defined over a set of random variables $x = \{x_i | i \in V\}$, where $V = \{1, 2, \ldots, n\}$. Each random variable x_i denotes a potential character in the word, and can take a label from the label set, $L = \{l_1, \ldots, l_k\} \cup \varepsilon$. The label set L is the set of English characters and digits, and a null label (ε) to suppress weak detections, similar to [1]. The most likely word represented by the set of characters x_i is found by minimizing the energy function, $E : L^n \rightarrow \mathbb{R}$, corresponding to the random field. The energy function $E(\cdot)$ can be typically written as sum of potential functions:

$$E(x) = \sum_{c \in \mathcal{C}} \psi_c(x_c), \tag{1}$$

where $\mathcal{C}$ represents a set of subsets of V, *i.e.* cliques, and x_c is the set of random variables included in a clique c. The set of potential characters is obtained by a sliding window based character detection step. The neighbourhood relations among characters, which determine the structure of the random field, are based on the spatial arrangement of characters in the word image. The character detection step provides us with a large set of windows potentially containing characters within them. Our goal is to infer the most likely word from this set of characters. We formulate this problem as that of minimizing the energy in (1), where the best energy solution represents the ground truth word we aim to find.

The energy function (1) is composed of unary, pairwise and higher order terms. The unary and pairwise terms are computed as described in [1]. For introducing higher order, we add an auxiliary variable x_c^a for every clique $c \in \mathcal{C}$. This auxiliary variable takes a label from the label set L_e. In our case the extended label set L_e, for a CRF of order h, contains all possible h-gram combinations present in the lexicons and one additional label (to account for h-grams that do not occur). We define a very high cost for an auxiliary variable to take a label which is not present in the dictionary. Increasing the order of the CRF allows us to capture a larger context. An illustration of our model is shown in Figure 1.

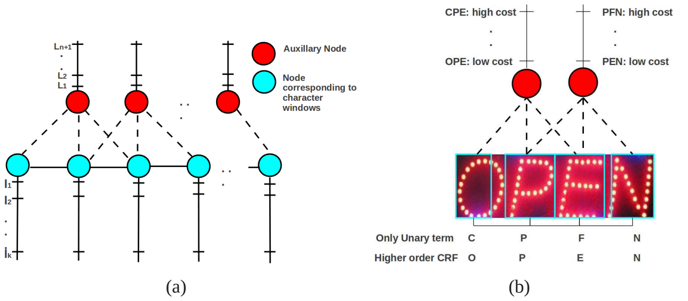

Figure 1: *The proposed graphical model (a model of order 3 is shown here) and an example word image to illustrate its use: Tri-grams like OPE, PEN have very high frequency in an English dictionary (> 1500), and thus are assigned a low cost, whereas unlikely tri-grams, such as CPE and PFN, are assigned a high cost.*

Method	SVT-WORD	ICDAR	IIIT 5K-word
Pairwise CRF [1]	23.49	45	20.25
Proposed Higher Order	**49.46**	**57.92**	**44.30**

Table 1: Word recognition accuracy without using an image specific small word list. Lexicon priors are computed from a large size lexicon with 0.5 million words.

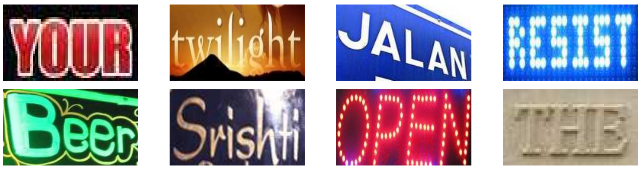

Figure 2: *A few sample images from the* IIIT *5K-word dataset where our method is successful. We see that the dataset contains images with variations in font, style, background, orientation etc.*

Results and Discussions. Our method outperforms [1] not only on the (smaller) SVT and ICDAR 2003 datasets, but also on the IIIT 5K-Word dataset[1]. We compare the word recognition performance of our method with pairwise CRF in Table 1. We achieve a significant improvement of around 25%, 12% and 22% on SVT, ICDAR 2003 and IIIT 5K-word datasets respectively. We also show few sample images from the 5K-word dataset in Figure 2.

Our method differs from other related approaches, such as [1], as detailed below. We address a more general problem of scene text recognition, *i.e.* recognizing a word without relying on a small size lexicon. Note that recent works [1, 2, 3] on scene text recognition, recognize a word with the help of an image-specific small size lexicon, of about 50 words per image. Our method computes the prior from an English dictionary and by-passes the use of edit distance based measures. In fact, we also recognize words missing from the given dictionary. One of the main reasons for the improvements we achieve is the use of n-grams extracted from the dictionary.

In summary, we proposed a powerful method to recognize scene text. The proposed CRF model infers the location of true characters, as well as the word as a whole. We evaluated our method on publicly available datasets and a large dataset introduced by us.

[1] A. Mishra, K. Alahari, and C. V. Jawahar. Top-down and bottom-up cues for scene text recognition. In *CVPR*, 2012.

[2] K. Wang and S. Belongie. Word spotting in the wild. In *ECCV*, 2010.

[3] K. Wang, B. Babenko, and S. Belongie. End-to-end scene text recognition. In *ICCV*, 2011.

[1] Our new dataset available at: http://cvit.iiit.ac.in/projects/SceneTextUnderstanding/