# Scenergy: Scene Text Recognition with Conditional Random Fields

Pranav Goyal*
*Dept. of Computer Science and Engineering*
*IIIT Delhi*
New Delhi, India
pranav17078@iiitd.ac.in

Surabhi S. Nath*
*Dept. of Computer Science and Engineering*
*IIIT Delhi*
New Delhi, India
surabhi16271@iiitd.ac.in

***Abstract*—Scene text recognition has been a popular subject of study for the computer vision community over the past ten years. More recent and novel systems try to incorporate the context into scene text recognition. We develop a method for sentence level text recognition in street signboard images, utilizing both top down and bottom up cues. A top down probabilistic graphical model works in synergy with the bottom up deep learning (DL) text recognition model. A conditional random field (CRF) is used to model the strength of relations across words in the image. An energy term consisting of weighted contributions from both the DL model and the CRF model is minimized to arrive at the optimal word sequence predictions. We analyze and compare the performance of 4 models, name the pure DL model, DL + linear CRF model, DL + dense fully connected CRF model and DL + dense CRF model with positional weighting. It is observed that the models with top down cues from the CRF exceed performance as compared to the pure DL model. Further, the linear CRF model performs better than the dense CRF model, while, the the dense CRF model with positional weighting performs better that the linear CRF model. This highlights the usefulness of lexicon derived bottom up cues and the role of positional effect on word sequence determination.**

***Index Terms*—Conditional Random Fields, Probabilistic Graphical Models, Scene Text Recognition**

## I. Introduction

### *Scene Text Recognition*

Scene text recognition has gained significant attention over the past few years. Most object detection [1] or semantic segmentation [2] strategies ignore the text information in the scene. Text plays a huge role in scene understanding as it captures attention and the contextual information [3]. Traditional OCR strategies work well on scanned texts [4], however they do not readily generalized to scenes with high variation in text fonts, colours, orientations and sizes [5]. Several techniques have emerged which can broadly be classified under three categories - Detection Methods, Recognition Methods and Auxiliary Methods. The Detection Methods aim to detect the positions of the text in the image while the Recognition Methods aim to determine the text in the detected regions. For Detection methods, multiple anchor-based [6], region-proposal based [7] methods have been developed. Sliding window methods in conjunction with CNNs [8] have been used to obtain potential text locations. Also, text instance level methods [9], connected-component based methods [5], and Bottom up techniques [10] have been used extensively. Further, the text is not restricted to straight snippets and can be oriented at different angles. Different methods have been developed to deal with multi-oriented text [11]. For recognition, attention-based methods [12], end-to-end encoder decoder models [13] and probabilistic models [14] have been used. The more modern auxiliary methods include the use of Synthetic Data [15], semi-supervised strategies [16], de-blurring [17], and other methods using contextual information [18], and adversarial techniques [19].

### *Conditional Random Fields*

The Conditional Random Field (CRF) is a discriminative undirected probabilistic graphical models which learns sequential data and takes context into account. The CRF models the conditional distribution and attains the optimal target sequence as follows:

$$\hat{y} = argmax_y p(Y|X) = argmax_y \frac{p(YX)}{p(X)}$$

where X is the set of observed nodes X and the Y is the sequences of targets. Further,

$$argmax_y \frac{p(YX)}{p(X)} = argmax_y (YX)$$

$$= argmax_y \sum_{i=1}^{n} -E(x_i, y_i) = argmin_y \sum_{i=1}^{n} E(x_i, y_i)$$

where n is the length of the word sequence and E(x, y) is the CRF energy expression

The CRF graph satisfies the property that when the graph is globally conditioned on X, the Ys satisfy the Markov property, that is, the depend only on their neighbours. Hence, this removes our need to explicitly model the dependencies between X's, as the dependency terms in the expression gets nullified when p(XY) is divided by p(X) yet the dependencies are inherently captured. Using this, minimizing the energy will yield the optimal word sequence ŷ.

CRFs have been commonly used for the purpose of

* equal contribution

scene text recognition. Pan et al. [20] used a CRF to remove non-text components from a potential set of sliding windows. Jaderberg et al. [21] proposed a CNN architecture with a CRF for recognition where the unary terms were obtained from a character level CNN and higher order terms were obtained from a n-gram level CNN. Zhang et al. [22] used proposed a two-step iterative CRF algorithm consisting of an OCR filtering step and a Belief Propagation inference step. Our works is based on Mishra et al.'s work on scene text recognition which exploits both bottom-up and top-down cues with the help of the CRF. Their work recognized character-level text and derived optimal word representations by minimizing the CRF energy. We extend this concept to detect word-level text and derive the optimal word sequence or sentence representations in a dataset of signboards images in real scenes.

## II. Methodology

The methodology is outlined in Figure 1. The bottom-up DL model and the top-down CRF model word together to output the most probable sequence of word predictions.

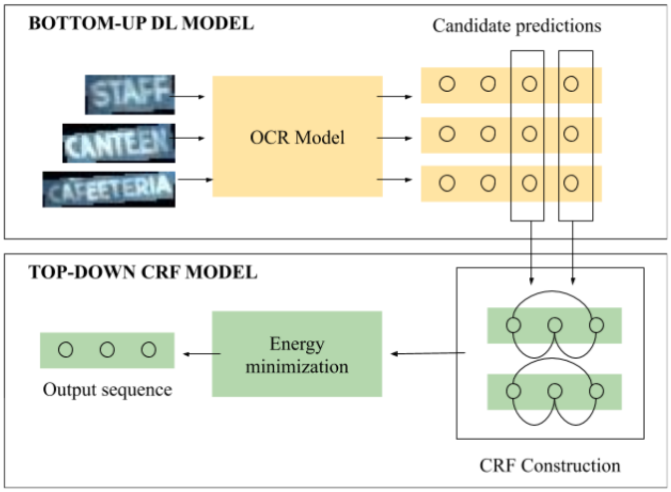

Fig. 1. DL and CRF models

### A. DL Model

We have used the model ASTER as our DL framework. The model has shown to produce near state of the art performance for scene text recognition on the various benchmark datasets. After obtaining a pre-trained ASTER trained on the MJSynth dataset containing 14 Million English words, we fine-tuned it on 8434 word images from our dataset. The model consists of three stages:

1) Transformation Network for converting the input word image into a more interpretable rectified image.
2) A BiLSTM encoder for feature extraction
3) An Attention based bidirectional decoder for obtaining character level probability values to then return the character combination with highest probability scores.

Since we wanted our outputs as a set of probable words, which could be fed to the CRF network for inference based on context, we modified the ASTER model output. Instead of returning the character combination with highest probability score, we returned a set of candidate words and corresponding probabilities, which included ASTER's prior output along with the lexicon words that lied within an edit distance of 3 from the prior output. These candidate predictions for each word were then fed to the Preprocessing Unit.

### B. Preprocessing Unit

In the preprocessing unit, the candidate predictions for each word obtained from the DL model are pruned and reduced to at most 3. This is achieved by taking the top 3 predictions with the highest unigram confidence value. These candidate predictions are then fed to the CRF Model.

### C. CRF Construction

Once the candidate predictions form the Preprocessing Unit are acquired for every word, combinations of these candidates are formed to obtain potential sequences and the energy for each sequence is calculated. The CRF Energy consists of a unigram term, obtained form the DL model, which represents the bottom-up cue, and a bigram term, obtained from the edge connections in the CRF. This acts as the top-down cue. These 2 terms are weighted by suitable constants $\lambda_u$ and $\lambda_b$ and added to obtain the energy of the sequence. If the candidate word pair is not found in the bigrams, an out of vocabulary penalty is added to the energy. Four different graphical constructions are tested:

1) Pure DL model: The $\lambda_b$ representing the bigram energy constant is set to 0
2) Linear CRF model: The bigram energy term is only considered by adjacent neighbours
3) Dense CRF model: The bigram term is considered assuming all words $y_i$ are connected to all other words $y_j$ where $i \neq j$
4) Dense CRF model with positional weighting: The bigram term is considered assuming all words $y_i$ are connected to all other words $y_j$ where $i \neq j$, however, the energies are assigned a weight if $\frac{1}{d}$, where d is the distance between the 2 words in the sequence

### D. Evaluation Metrics

We have tested, analysed and compared the performance of the above four mentioned models on 3 different metrics, namely word match accuracy, average edit distance and average phonetic distance.

*1) Word Match Accuracy:* Word match accuracy is obtained by calculating the number of correct word matches, divided by the total number of words in the test set.

$$WM = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} \mathbb{I}(\hat{y}_{ij}, y_{ij})}{\sum_{i=1}^{m} \sum_{j=1}^{n_i} n_i}$$

where m is the number of images, $n_i$ is the number of words in image i and $\mathbb{I}(\hat{y}_{ij}, y_{ij}) = 1$ if $\hat{y}_{ij} = y_{ij}$, otherwise 0

*2) Average Edit Distance:* Average edit distance is obtained by calculating the mean of the total edit distance between actual and predicted sequence across all images in the test set. This was incorporated to prevent penalizing minor word differences, which potentially yield in character level similarity.

$$AED = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} ED(\hat{y}_{ij}, y_{ij})}{m}$$

where m is the number of images, $n_i$ is the number of words in image i and $ED(\hat{y}_{ij}, y_{ij})$ is the edit distance between $\hat{y}_{ij}$ and $y_{ij}$ calculated using Dynamic Programming

*3) Average Phonetic Distance:* Average phonetic distance is obtained by calculating the mean of the total edit distance between actual and predicted sequence across all images in the test set. This was incorporated to prevent penalizing minor word differences, which potentially yield in similar pronunciation sound.

$$APD = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} PD(\hat{y}_{ij}, y_{ij})}{m}$$

where m is the number of images, $n_i$ is the number of words in image i and $PD(\hat{y}_{ij}, y_{ij})$ is the phonetic distance between $\hat{y}_{ij}$ and $y_{ij}$ calculated as a weighted combination of phonetic key outputs from 3 different algorithms - Soundex, NYSISS and Metaphone, in the ratio 0.3:0.2:0.5.

## III. Experiments and Results

### A. Dataset

The dataset consists of 1300 street signboard images for the task of scene text recognition. Each signboard consists of multiple words in Hindi and English. The Hindi words were not considered for our study. The text is multi-oriented and images were shot from different angles and readability conditions. Each signboard is associated with coordinates of words and the text in these images for the tasks of recognition. We extracted 9044 English word cropped images using the ground truth coordinates for recognition, while retaining the information of which signboards they belonged to. Sample images are shown in Figure 2. The data was split into train and test sets. The train set comprised of 1240 signboard images, with 8434 word images, while the test set comprised of 60 signboard images with 610 word images. Data statistics for the test set is shown in Figure 3. The bigram probabilities represented the joint probability of co-occurrence probability of 2 words. The bigram value for (a, b) was obtained by calculating the number of signboards with both a and b, divided by the total number of signboard images.

### B. Experiments

A series of experiments were performed comparing and contrasting the performance of the 4 models on the train and test data. The DL model was trained fine-tuned for 7 epochs. The train and test set loss and accuracy plots are shown in Figure 4. The train and test candidates were then preprocessed and passed to the CRF block. Energy minimization was performed for all 4 models followed by evaluation.

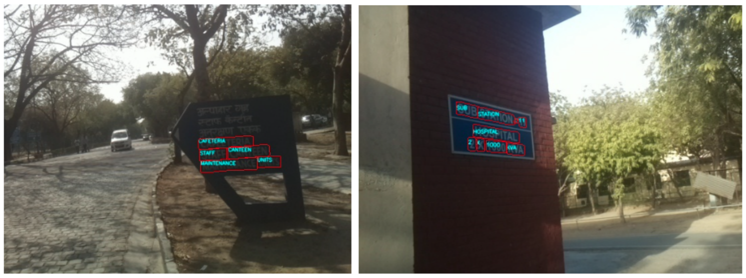

Fig. 2. Sample dataset images with ground truth

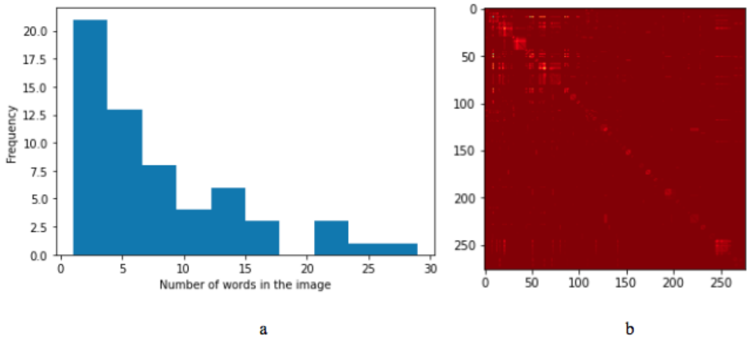

Fig. 3. Data Statistics for the test set: a) histogram of number of words in an image, b) heatmap of the bigram probabilities

### C. Results

The results indicate that the CRF top down cues were able to improve performance as compared to the pure DL model on both train and test set. Further, on the test set, the linear CRF model performs better than the dense CRF model, while, the the dense CRF model with positional weighting performs better that the linear CRF model. On the train set, the performance of dense CRF with and without positional weighting marginally exceeds performance as compared to the linear CRF. This highlights the usefulness of lexicon derived bottom up cues and the role of positional effect on word sequence determination. The results are shown in Figure 5.

## IV. Observations

We have some interesting observations from our experiments. The results show that the top down cues from the CRF are valuable for sequence determination. We observe around 3% increase on the train set and up to 3.63% increase in word match accuracy on the test set. This increase can be explained as the CRF is able to capture the interactions between the lexicon words and use that to aid sequence determination. In other words, this means that simply choosing the DL model output with highest confidence (lowest energy), may not always be optimal. The correct word may belong to the candidate set but may be predicted with lower confidence, in-turn may have higher unary energy. The bigram energy allows the sequence to incorporate such words in the sequence, thus improving performance. It should

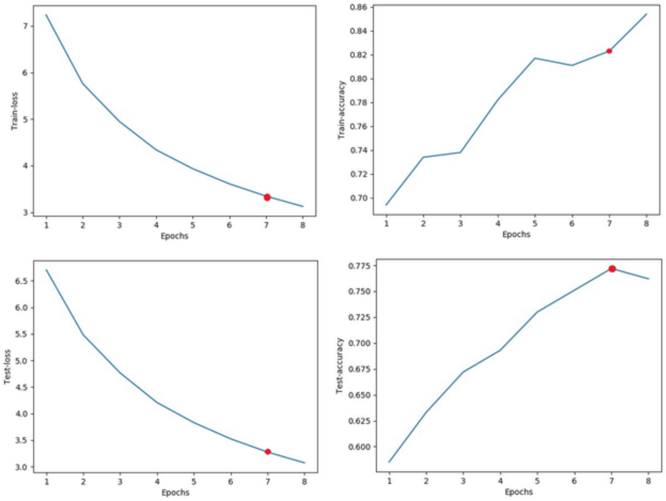

Fig. 4. Data Statistics for the test set

| | | Train Set | Test Set |
|---|---|---|---|
| Pure DL Model | Word Match Accuracy | 81.88 | 76.73 |
| | Average Edit Distance | 5.33 | 6.24 |
| | Average Phonetic Distance | 2.69 | 3.22 |
| Linear CRF | Word Match Accuracy | 84.07 | 80.06 |
| | Average Edit Distance | 5.19 | 6.11 |
| | Average Phonetic Distance | 2.63 | 3.19 |
| Dense CRF | Word Match Accuracy | 84.70 | 78.24 |
| | Average Edit Distance | 5.14 | 6.35 |
| | Average Phonetic Distance | 2.59 | 3.35 |
| Dense CRF with Positional Weighting | Word Match Accuracy | 84.65 | 80.36 |
| | Average Edit Distance | 5.12 | 5.94 |
| | Average Phonetic Distance | 2.59 | 3.10 |

Fig. 5. Results on the 4 models

be pointed out here that the top down cues will be useful only when majority of the words in the test set images have appeared in the train set image words. If not, the bigram energy would only end up penalizing the out of vocabulary words.

The dense CRF uses bigram contributions from every pair of words in the candidate sequence, while the linear CRF uses contributions only from the adjacent neighbouring words as per the signboard. Thus, the linear CRF captures local properties, while the dense CRF captures global signboard level properties.

It is seen that the Linear CRF tends to perform better than the Dense CRF on the test set, although Dense CRF with positional weights performs better than the Linear CRF. On the train set however, both dense CRF models have similar performance, around 0.6% better than the linear CRF. This can be explained as follows. The dense CRF has an upper hand when the signboard images are repeated, as in the case of different angles and illuminations images of the same signboard, as it learns signboard level word relations. Such cases of multiple images of the same signboard are more prevalent in the train set. Two sample images are shown in Figure 6. Thus, dense CRFs tend to perform better as compared to linear CRF in the train set.

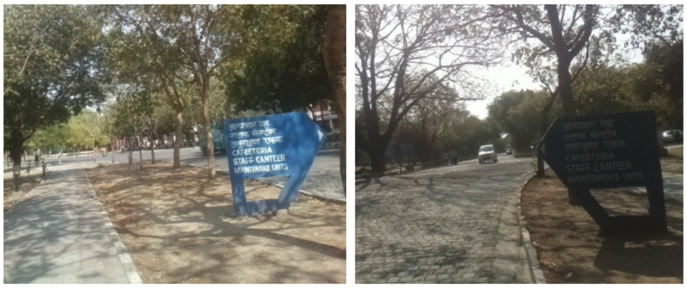

Fig. 6. Examples where dense CRF works better

In the test set however, the ability to correctly determine local phrases drives performances. It is observed that most locations have either 2 or 3 word length name, such as ”Taxila Apartments”, ”Maintenance Unit”, ”Student Activity Centre”, etc. It is often seen that the same word pair occurs in 2 signboards with different overall text content. Sample images are shown in Figure 7. In such cases, the dense CRF is not able to perform at par with the linear or dense positional weighted CRF due to lack of the ability to prioritize local word correlations. The linear CRF is a special case of the dense CRF with positional weights, where all not adjacent contributions are 0, while the positional weighted CRF evenly reduces contribution weight with increasing distance between the candidate words. These 2 models are able to give more importance to adjacent word contributions. Hence, they can learn local features and work well on data in the test set.

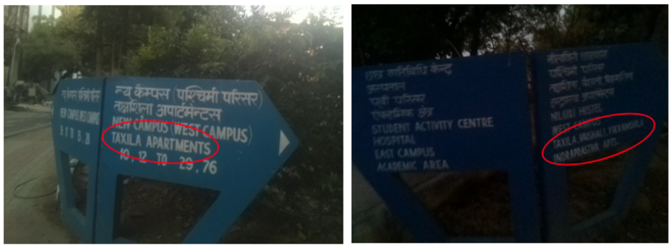

Fig. 7. Examples where dense CRF works better

## V. Individual Contribution

We worked in a symbiotic manner, in full *”scenergy”*. We divided our work equally and efficiently between the bottom-up backend DL model, and the top-down front-end CRF model components of our project. Pranav worked on the DL model, including training, fine-tuning and testing on the DL model, obtaining the candidate predictions and bigram probabilities. Surabhi worked on the front end, including CRF construction, preprocessing, energy minimization and parameter estimation.

# References

[1] C. Desai, D. Ramanan, and C. Fowlkes, ”Discriminative models for multi-class object layout”, *ICCV*, 2009.

[2] J. Shotton, J. Winn, C. Rother, and A. Criminisi, ”Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context”, *IJCV*, 81(1):2–23, 2009.

[3] T. Judd, K. Ehinger, F. Durand, and A. Torralba, ”Learning to predict where humans look”, *ICCV*, 2009.

[4] Uchida S, ”Text localization and recognition in images and video”, *Handbook of document image processing and recognition*, Springer, London, pp 843–883, 2014.

[5] H. Lin, P. Yang and F. Zhang, ”Review of Scene Text Detection and Recognition”, *Archives of Computational Methods in Engineering*, 2020.

[6] Liu, Yuliang and Jin, Lianwen, ”Deep Matching Prior Network: Toward Tighter Multi-oriented Text Detection”, CVPR, 2017.

[7] Liao, Minghui, Zhu, Zhen, Shi, Baoguang, Xia, Gui-song and Bai, Xiang, ”Rotation-Sensitive Regression for Oriented Scene Text Detection”, *CVPR*, 2018.

[8] Wang T, Wu DJ, Coates A, Ng AY, ”End-to-end text recognition with convolutional neural networks”, *International conference on pattern recognition*, pp 3304–3308, 2012.

[9] Liao, Minghui, Shi, Baoguang, Bai, Xiang, Wang, Xinggang and Liu, Wenyu, ”TextBoxes: A Fast Text Detector with a Single Deep Neural Network”, *AAAI*, 2017.

[10] Dan, Deng, Haifeng, Liu, Xuelong, Li and Deng, Cai, ”PixelLink: Detecting Scene Text via Instance Segmentation”, *AAAI*, 2018.

[11] Shi, Baoguang, Bai, Xiang and Belongie, Serge, ”Detecting Oriented Text in Natural Images by Linking Segments”, *CVPR*, 2017.

[12] Ghosh, Suman K, Valveny, Ernest and Bagdanov, Andrew D, ”Visual attention models for scene text recognition”, *ICDAR*, 2017.

[13] Busta, Michal, Neumann, Lukas and Matas, Jiri, ”Deep TextSpotter: An End-To-End Trainable Scene Text Localization and Recognition Framework”, *ICCV*, 2017.

[14] Jerod J Weinman, ”Scene Text Recognition Using Similarity and a Lexicon with Sparse Belief Propagation”, *TPAMI*, 2009.

[15] Jaderberg, Max, Simonyan, Karen, Vedaldi, Andrea and Zisserman, Andrew, ”Synthetic data and artificial neural networks for natural scene text recognition”, *NIPS*, 2014.

[16] Tian, Shangxuan, Lu, Shijian and Li, Chongshou, ”Wetext: Scene text detection under weak supervision”, *ICCV*, 2017.

[17] Hradis, Michal, Kotera, Jan, Zemcik, Pavel and Sroubek, Filip, ”Convolutional neural networks for direct text deblurring”, *BMVC*, 2015.

[18] Zhu, Anna, Gao, Renwu and Uchida, Seiichi, ”Could scene context be beneficial for scene text detection?” *Pattern Recognition*, 2016.

[19] Yuan, Xiaoyong, He, Pan and Li, Xiaolin Andy, ”Adaptive Adversarial Attack on Scene Text Recognition”, *CVPR*, 2018.

[20] Pan YF, Hou X, Liu CL, ”A hybrid approach to detect and localize texts in natural scene images”, *IEEE Transactions on Image Processing*, 20:800–813, 2011.

[21] Jaderberg M, Simonyan K, Vedaldi A, Zisserman A, ”Deep structured output learning for unconstrained text recognition”, *International conference on learning representations*, pp 1–10, 2015.

[22] H. Zhang, C. Liu, C. Yang, X. Ding and K. Wang. ”An Improved Scene Text Extraction Method Using Conditional Random Field and Optical Character Recognition”. *International Conference on Document Analysis and Recognition*, 2011.