Top-down and Bottom-up Cues for Scene Text Recognition

Anand Mishra[1], Karteek Alahari[2] and C.V. Jawahar[1]

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

Text is important
- Information rich
- Useful cues
- Viewers fixate on text more [ICCV'09]

Text Detection
- Adaboost based [CVPR'04, ICDAR'11]
- SWT (CVPR'09)
- Multi-oriented text detection [CVPR'12]

Word Recognition
- IP based [CVPR'11]
- Sparse BP [TPAMI'09]
- PLEX and PICT [ECCV'10, ICCV'11]

Detection and Recognition
- Real time text localization and recognition [CVPR'12]
- PLEX [ICCV'11]

Many Applications
- Multi-media indexing
- Mobile apps
- Auto navigation
- Help for visually impaired

The Goal

Recognize a cropped word

Datasets

Sign Evaluation [Weinman *et al.* PAMI'09]

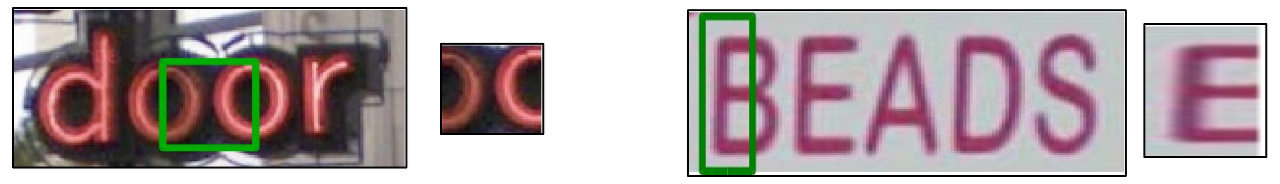
ICDAR 2003

Street View Text [Wang *et al.*, ECCV'10]

- State-of-the-art commercial OCR : low accuracy
- Sign Evaluation (60.5%), ICDAR (56%), Street View Text (35%)

Challenges

- Inter and intra character confusion

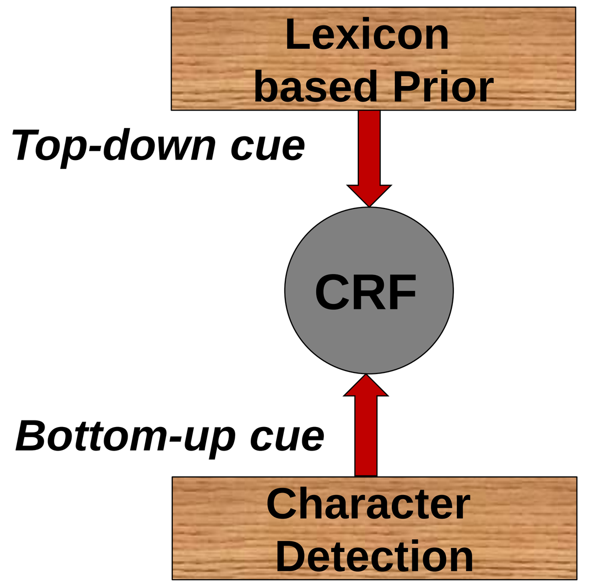


- Large number of classes
- Poor isolated character recognition

Need strong cues

Top-down and Bottom-up cues

- **Top-Down:** Prior computed from lexicon
- **Bottom-up:** Sliding window based character detections
- The CRF model infers the true characters and the word as a whole.



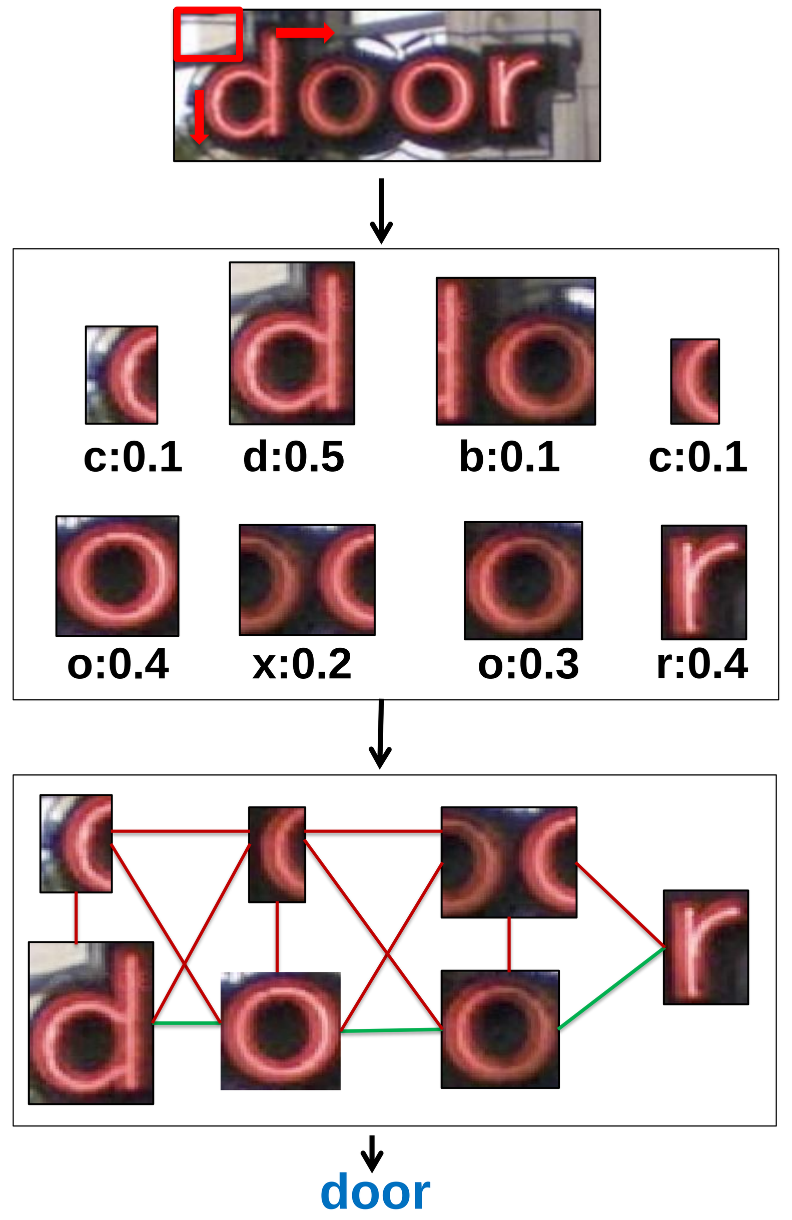
Method Overview

Character detection

- Sliding window
- SVM classifier trained on ICDAR'03
- HoG features
- Some windows are pruned based on aspect ratio

Graph construction

- Character windows = nodes
- Unary cost = $1 - f(SVM\ Score)$
- Pairwise cost = Lexicon + overlap based



The CRF Energy

- Set of labels $L = \{0, 1, \ldots, 9, a, b, \ldots z, A, B, \ldots, Z, \epsilon\}$
- Minimize an energy of following form:

$$E(X) = \sum_{i=1}^{n} E_i(x_i) + \sum_{\mathcal{E}} E_{ij}(x_i, x_j)$$

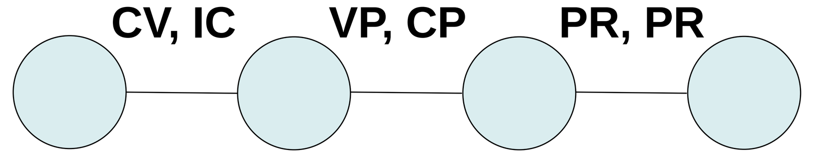
Unary cost: $E_i(x_i = c_j) = 1 - P(c_j|x_i)$

Unary cost of ϵ is computed from SVM score and aspect ratio prior.

Pairwise cost:
- Lexicon based: $E_{ij}(x_i = c_i, x_j = c_j) = \lambda_l(1 - P(c_i, c_j))$
- Overlap based: $E_{ij}(x_i = c_i, x_j = c_j) = \lambda_o exp(-(100 - overlap(x_i, x_j))$

Prior Computation

Toy example: Bi-gram prior v/s node specific prior



P(CV)	1/6	1/2	1/6	0	1/6	0
P(IC)	1/6	1/2	1/6	0	1/6	0
P(VP)	1/6	0	1/6	1/2	1/6	0
P(CP)	1/6	0	1/6	1/2	1/6	0
P(PR)	1/3	0	1/3	0	1/3	1

Lexicon = {CVPR, ICPR}

Possible character pairs = {CV, IC, VP, CP, PR,PR}

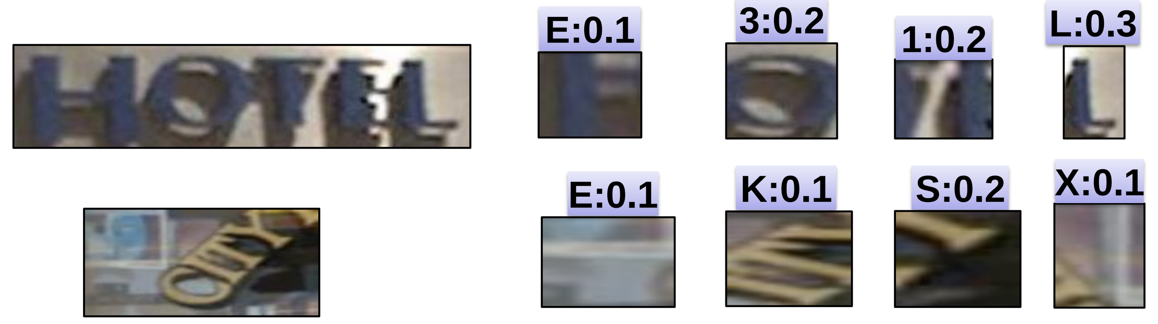
Implementation Details

- **Descriptor:** Dense HOG, cell size = 4 × 4, bins = 10 bins, after resizing image to a 22× 20.
- **Inference:** Tree-reweighted message passing (TRW-S) [Kolmogorov, TPAMI'06].
- The method is as it is applicable to near frontal text datasets like Sign Evaluation data too.

Results

Method	SVT-Word	ICDAR(50)
PICT	59	-
PLEX+ICDAR	56	72
ABBYY 9.0	35	56
Proposed (bi-gram)	70.03	76.96
Proposed (node specific)	73.26	81.78

Failure cases



Weak detection and low SVM scores

Summary

- A general framework for scene text recognition
- Improves accuracies significantly on ICDAR and SVT
- Joint Probabilistic inference with lexicon priors unlike [ICCV'11]
- The method deals with poor character detections unlike [TPAMI'09]

Center for Visual Information Technology

International Institute of Information Technology, Hyderabad, INDIA

http://cvit.iiit.ac.in/projects/SceneTextUnderstanding

Travel grant supported by Google Inc. India