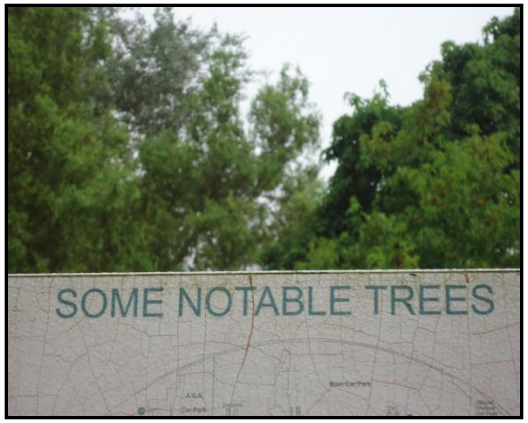
Scene Text Recognition using Higher Order Language Priors

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

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

Natural Scene Text: Why?



Text is everywhere !!

Many fundamental problems: Detection, Segmentation and Recognition

Many applications: mobile apps, auto navigations, multi-media indexing etc.

Natural Scene Text: Recent Interest

Detecting Text in Image

Stroke Width Transform based text detection [Epshtein *et al.*, CVPR'10]

End-to-end Scene Text Recognition [Wang and Belongie, ICCV' 11]

Real time localization and recognition [Neumann and Matas, CVPR'12]

Natural Scene Text: Recent Interest

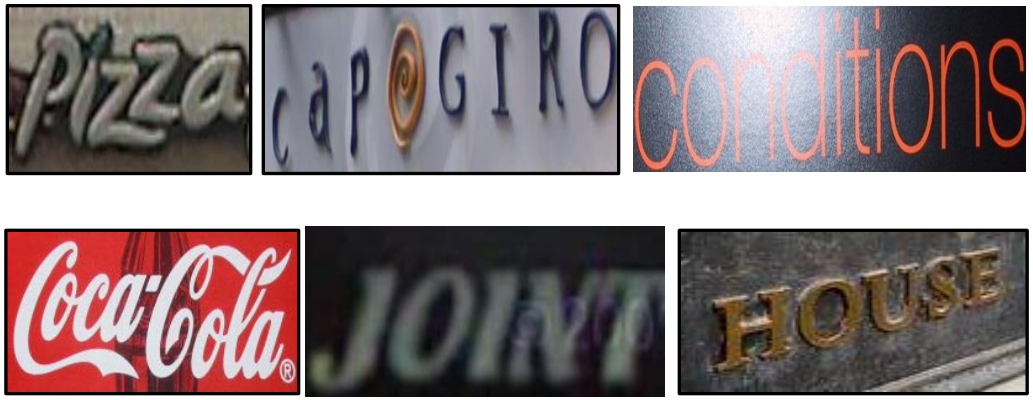
Text Recognition

Exemplar Driven Character Recognition in the Wild [Sheshadri and Divvala, BMVC'12]

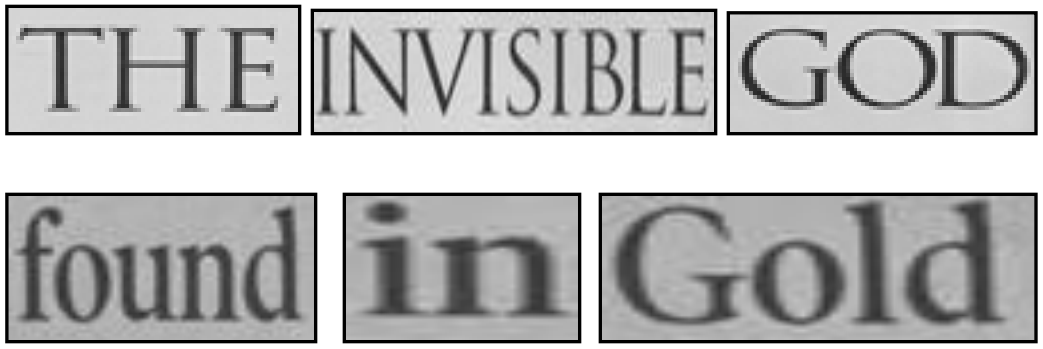
PLEX and PICT [Wang and Belongie, ECCV'10, ICCV'11]

Top-down and Bottom-up cues [Mishra *et al.*, CVPR'12]

Scene Text Recognition

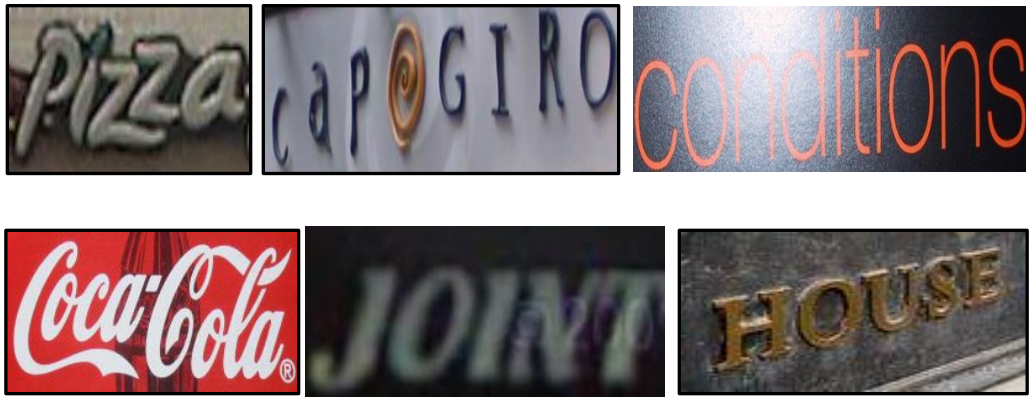


v/s

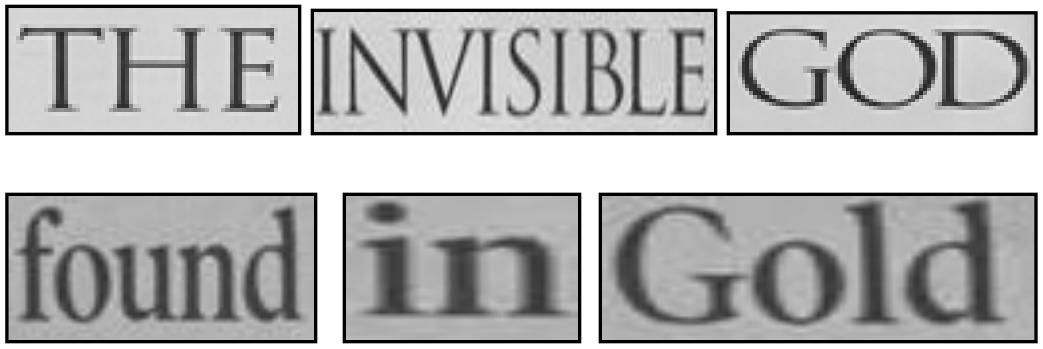


- Scene Text Recognition ≠ Optical Character Recognition (OCR)
- Good segmentation is tough
- Isolated character recognition accuracies are very low
- Not practical as a Classification problem

Scene Text Recognition



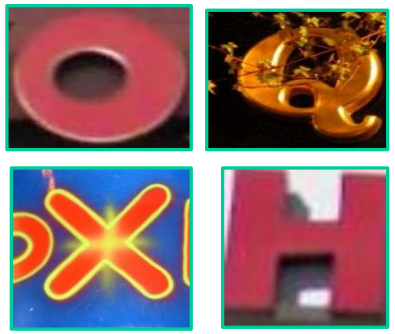
v/s



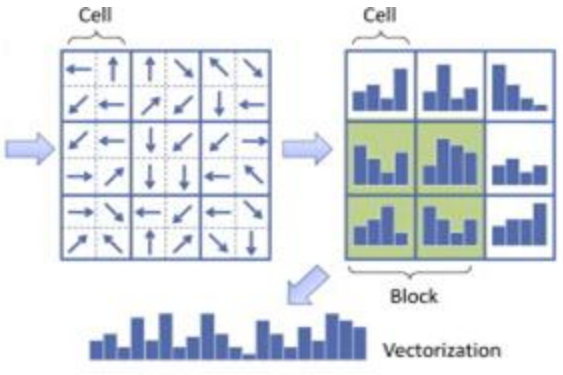
- Scene Text Recognition ≠ Optical Character Recognition (OCR)
- Good segmentation is tough
- Isolated character recognition accuracies are very low
- Not practical as a Classification problem

We need a better model

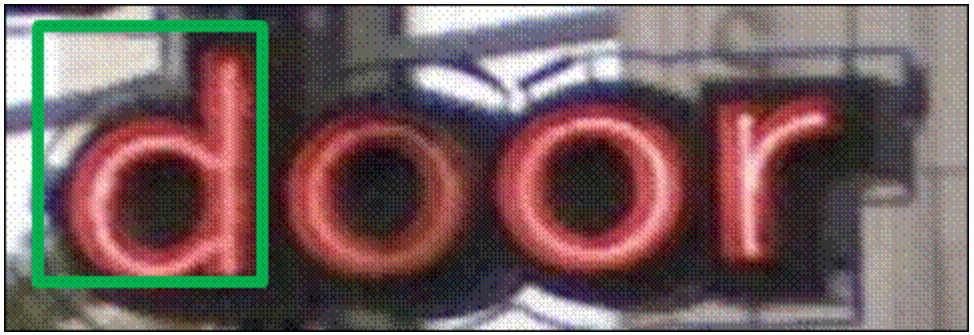
The Probabilistic Model



Character training data

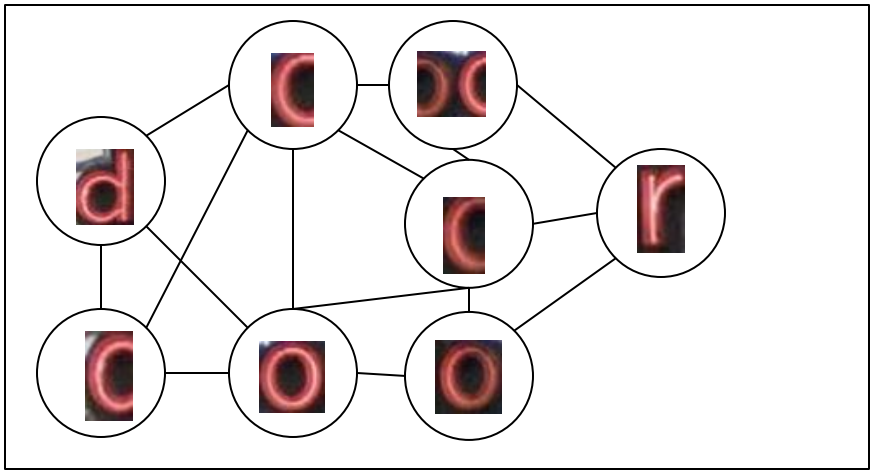


SVM



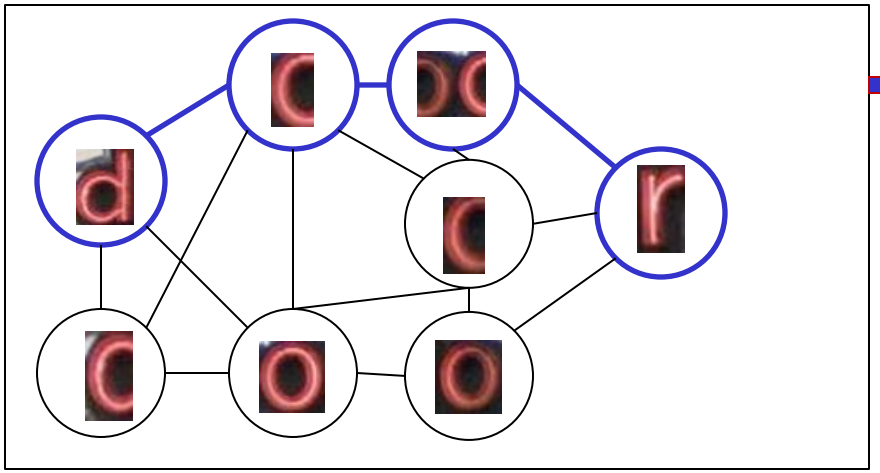
- Do not rely on “hard” segmentation
- HoG features
- Multi-class SVM trained on character level
- Sliding window based character detection

The Probabilistic Model



The Probabilistic Model

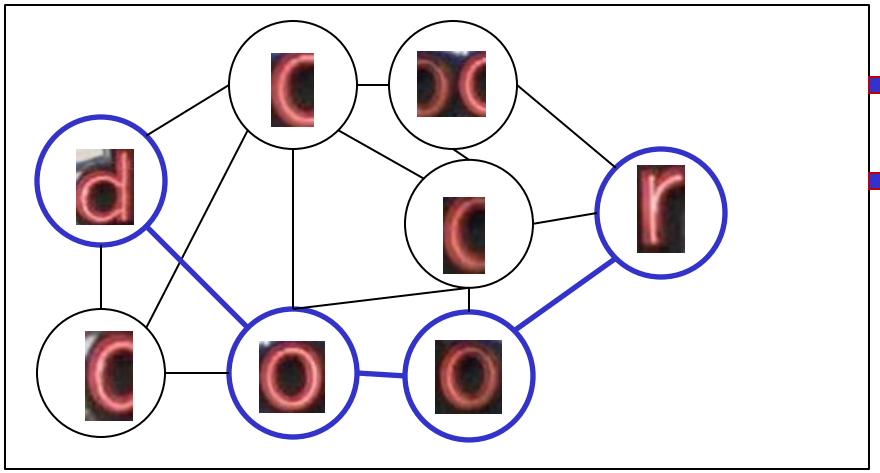
$$L = \{0, 1, \ldots, 9, a, b, \ldots z, A, B, \ldots, Z, \epsilon\}$$



dcxr

The Probabilistic Model

$$L = \{0, 1, \ldots, 9, a, b, \ldots z, A, B, \ldots, Z, \epsilon\}$$

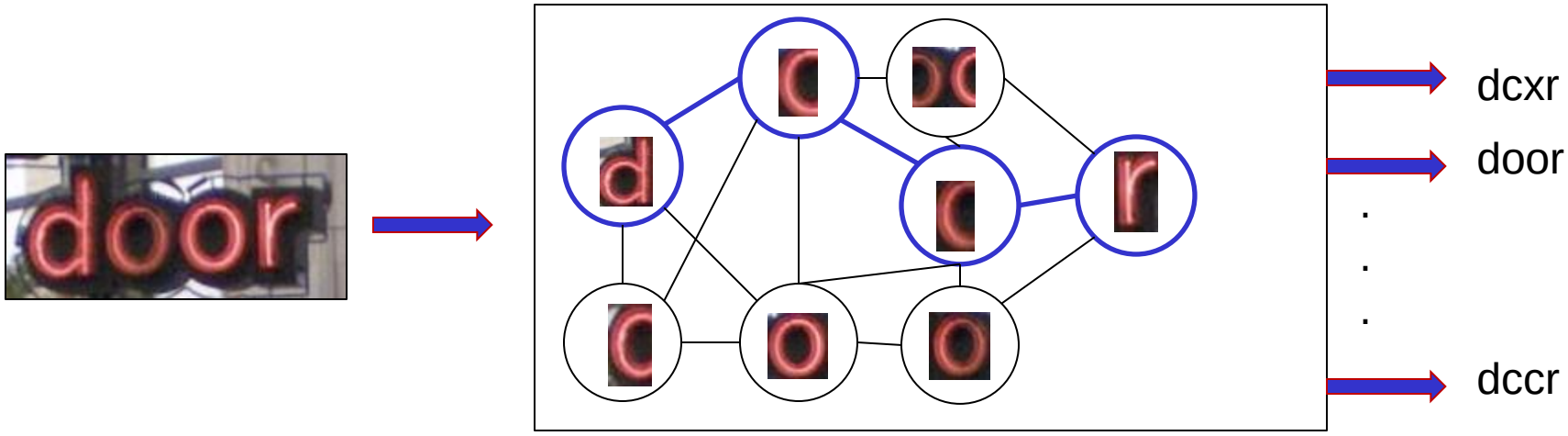


dcxr

door

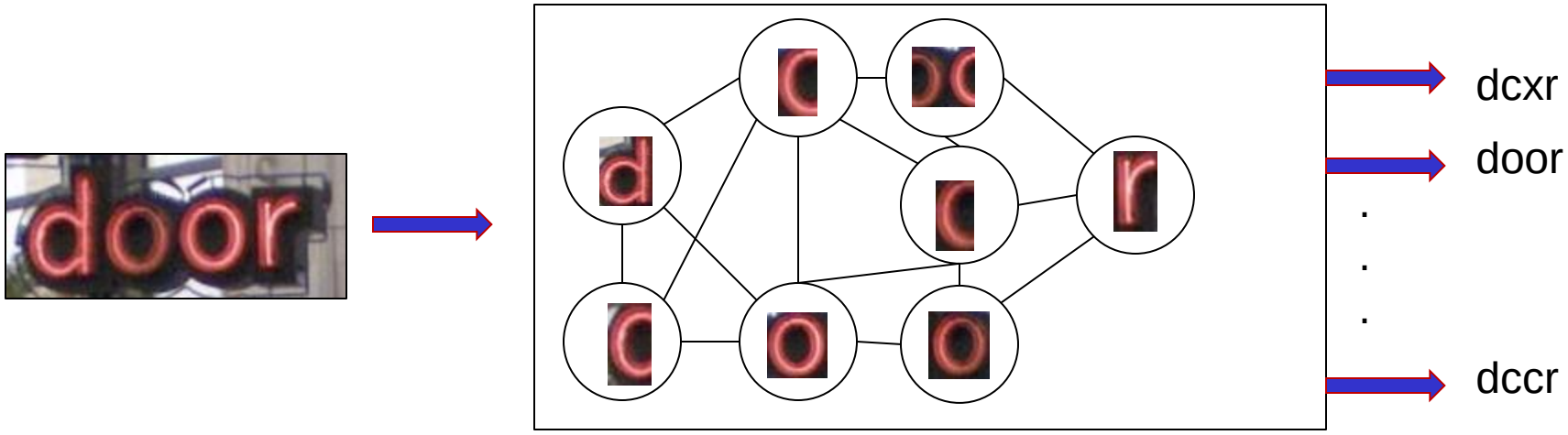
The Probabilistic Model

$$L = \{0, 1, \ldots, 9, a, b, \ldots z, A, B, \ldots, Z, \epsilon\}$$



The Probabilistic Model

$$\boldsymbol{L} = \{0, 1, \ldots, 9, \mathrm{a}, \mathrm{b}, \ldots \mathrm{z}, \mathrm{A}, \mathrm{B}, \ldots, \mathrm{Z}, \epsilon\}$$



- Many-many labelings possible, which is the optimal?
- In general, the problem is NP-Hard
- We solve the approximate version of the problem in an energy minimization framework.

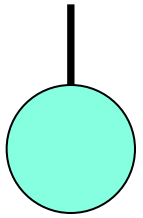
$$E(x) = E_i(x_i) + E_{ij}(x_i, x_j) + E_{ij\ldots p}(x_i, x_j, \ldots, x_p)$$

The CRF Energy

- The unary term

SVM score for character class c_j

$$E_i(x_i = c_j) = 1 - P(c_j|x_i)$$

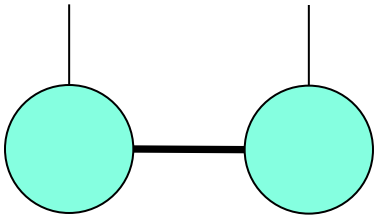


- The pair-wise term

Lexicon based:

$$E_{ij}(x_i = c_i, x_j = c_j) = \lambda_l(1 - P(c_i, c_j))$$

Joint probability for character class c_i and c_j



Overlap based:

$$E_{ij}(x_i, x_j) = \lambda_o exp(-(100 - overlap(x_i, x_j))$$

The cost of two highly overlapping nodes taking non-null label is high

The CRF Energy

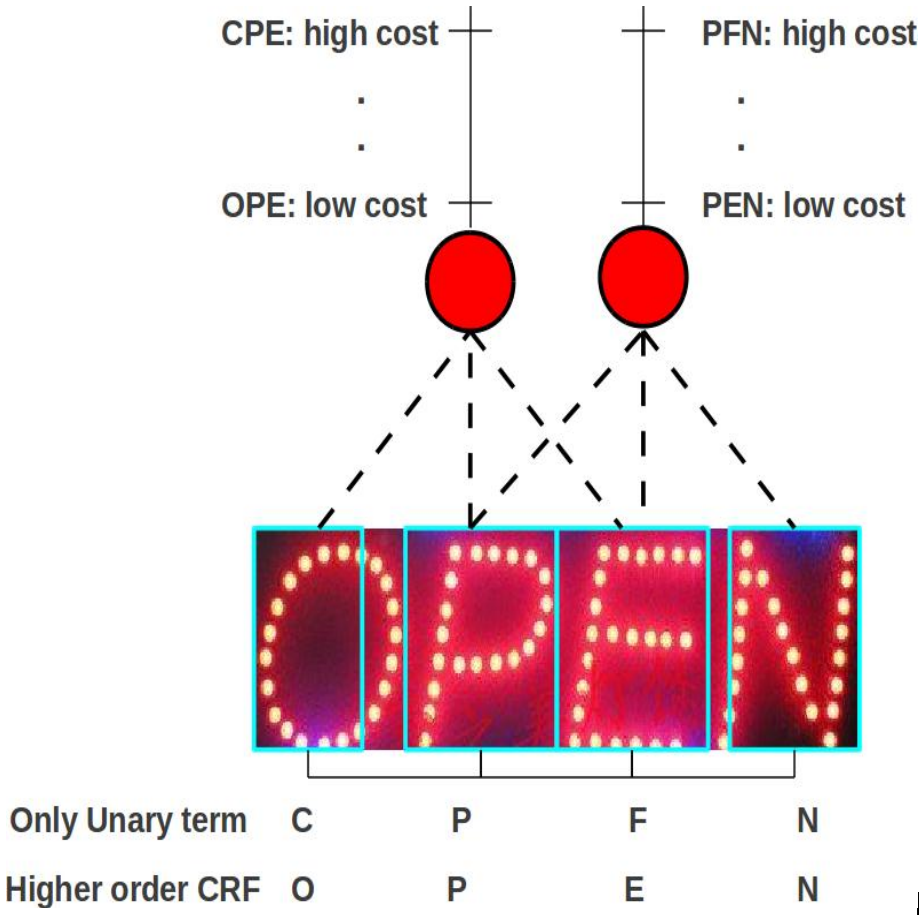
- Higher Order term
  = Unary + Pairwise

- Joint probability space of character sets occurring together is sparse

- Unary cost

$$E^a(x_i = L_i) = \lambda_a(1 - p(L_i))$$

- Pairwise cost to prevent non-dictionary *n-grams*



Prior Computation

Pair-wise Priors

- Bi-gram Priors
  - Joint Probability

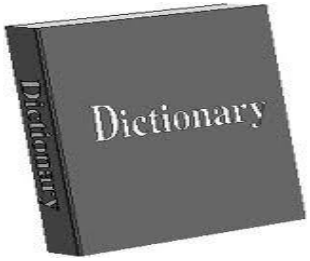
$P(c_i, c_j)$

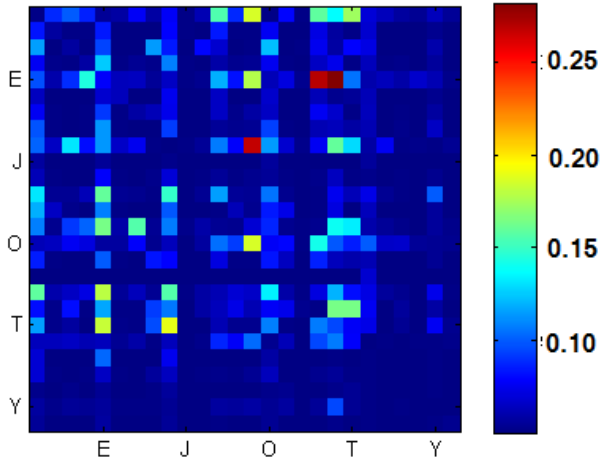
- Node Specific Priors
  - Joint Probability based on spatial position

Higher Order Priors

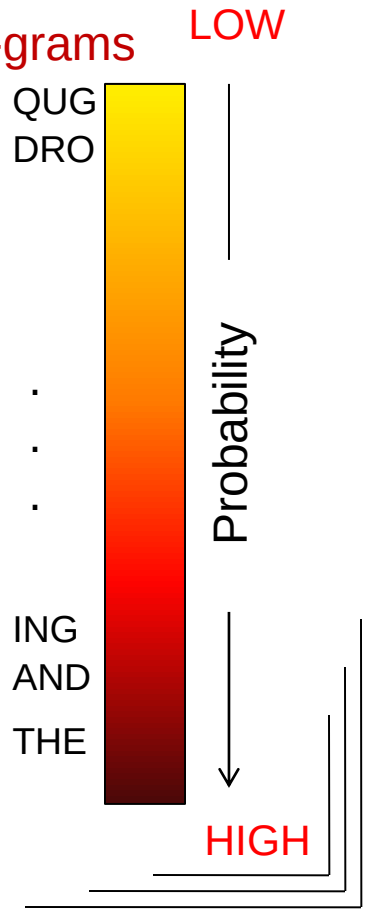
- n-gram Priors
  - Joint Probability

$P(c_i, c_j, \ldots, c_n)$

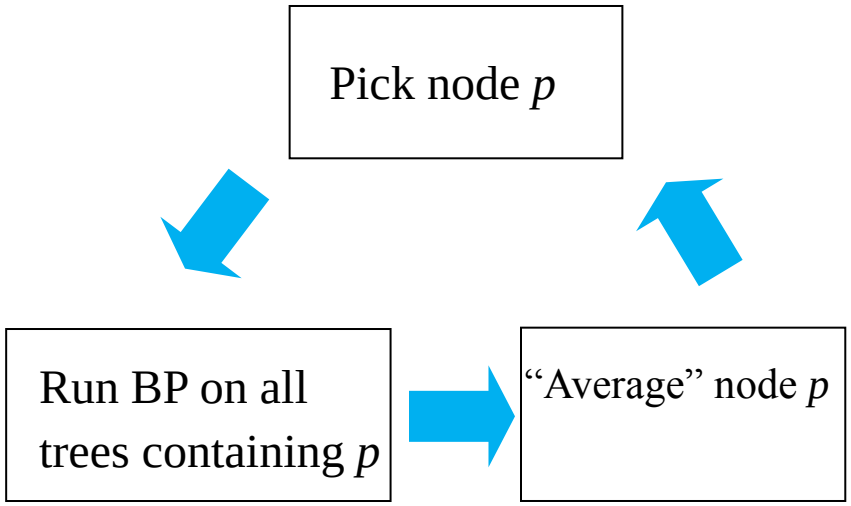




Bigram Probability distribution



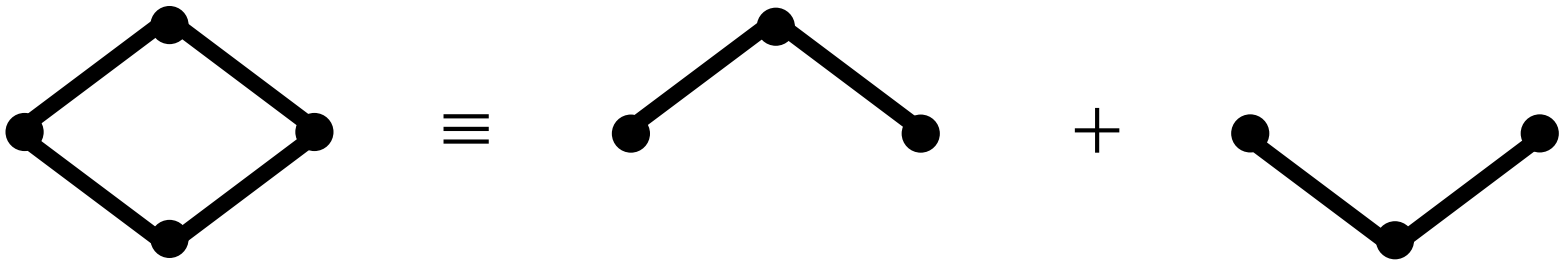
Inference



- Minimize energy of following form:

$$E(x) = E_i(x_i) + E_{ij}(x_i, x_j) + E_{ij\ldots p}(x_i, x_j, \ldots, x_p)$$

- We use Tree Re-weighted Message Passing (TRW-S) to minimize the energy



[$] Kolmogorov, **Convergent Tree-Reweighted Message Passing for Energy Minimization**,TPAMI'06

Lexicon driven v/s Lexicon free Recognition

Recognize a cropped word

Lexicons

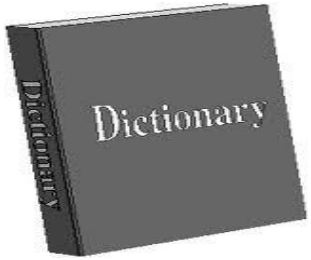
CAPOGIRO

Lexicons = Grocery item list

- Many Applications
  - e.g. assisting visually impaired person to navigate in a Grocery store

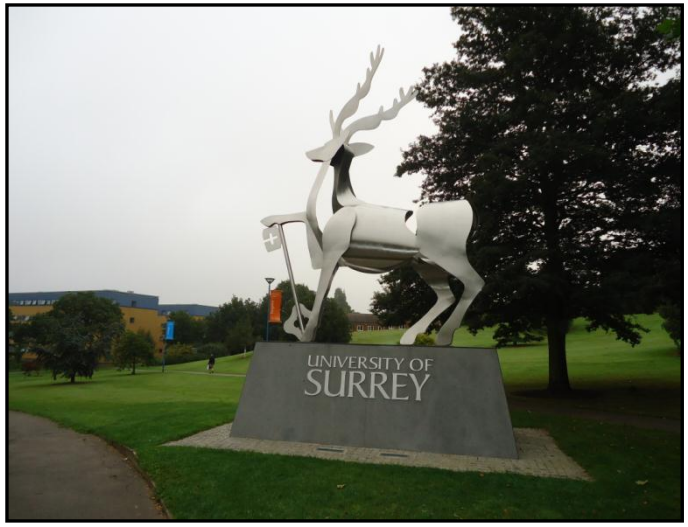
Lexicon driven v/s Lexicon free Recognition

Recognize a cropped word



CAPOGIRO

- Many Applications
  - e.g. unconstrained word recognition
  - Word may or may not belong to dictionary



IIIT 5K-word dataset

- 5000 words: Street View and born-digital images
- At-least 6 times large than popular public datasets
- Wide Variations
- *Up to character boundary level annotation*

IIIT 5K-word dataset

- Train set: 2000 words, Test set: 3000 words
- Collected from total 1120 scene/born-digital images
- *Grouped into easy/hard*

Lexicon driven Recognition

Method	SVT-Word	ICDAR(50)	IIIT 5K word
PICT [ECCV'10]	59	-	-
PLEX[ICCV'11]	57	72	-
ABBYY 9.0	35	56	14.60
Proposed[$] (Pair-wise)	**73.26**	**81.78**	**66**
Proposed (Higher Order)	**73.27**	**80.28**	**68.25**

Smaller lexicon: stronger context

Pairwise priors are powerful

Edit distance based corrections are possible

[$] [Mishra *et al.*, CVPR 2012]

Lexicon free Recognition

Datasets	ABBYY9.0	Pair-wise	Higher Order(=4)
SVT-Word	32.6	23.49	**49.46**
ICDAR2003	52	45	**57.92**
IIIT 5K-Word	14.60	20.25	**43.3**

- These experiments do not use any edit distance based correction
- 0.5 Million dictionary words are used to compute priors
- ICDAR2003 words with special characters are avoided.

Qualitative Results

	Unary	Pair-wise	Higher Order
	Y0UK	YOUK	YOUR
	TWI1IOHT	TWILIOHT	TWILIGHT
	KE5I5T	KESIST	RESIST
	DOLCE	DOLCE	DOLCE
	BEE1	BEEI	BEER
	SRISNTI	SRISNTI	SRISHTI

Conclusions and On-going Work

- A principled framework
- Joint inference to detect true characters and recognize word as a whole
- A novel higher order potentials
- Largest dataset for Scene Text Recognition

Ongoing Work

- Better features and learning for scene character classification

Thank You

Visit our project page for more detail of the work
http://cvit.iiit.ac.in/projects/SceneTextUnderstanding/

Supported by Microsoft Research India PhD Fellowship