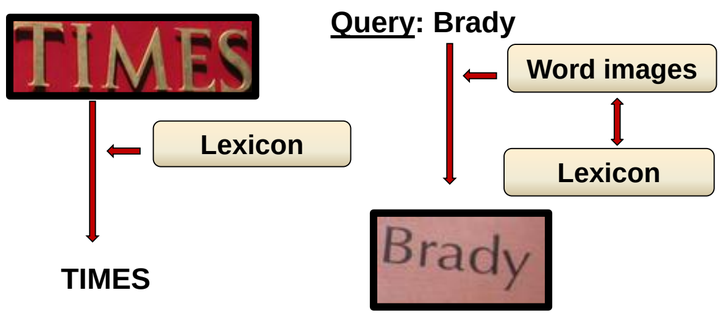
Scene Text Recognition and Retrieval for Large Lexicons

Udit Roy[1], Anand Mishra[1], Karteek Alahari[2] and C. V. Jawahar[1]
*[1]IIIT Hyderabad, India [2]Inria**

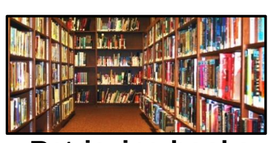
Goal

Word Recognition **Text-to-Image Retrieval**



Some Applications

Recognizing grocery items



Retrieving books by their titles

Our Approach

- Focus on large lexicon based recognition
- Multiple candidate words generation
- Inferring diverse solutions
- Group edit distance based lexicon re-ranking
- Iterative lexicon reduction
- Text-to-image retrieval task
  - Preprocess images by reducing the lexicon
  - Retrieve word images with the query word in the reduced lexicon

Main Idea:

Strengthen pairwise terms by reducing lexicon size using diverse solutions

Previous Work

Lexicon driven recognition

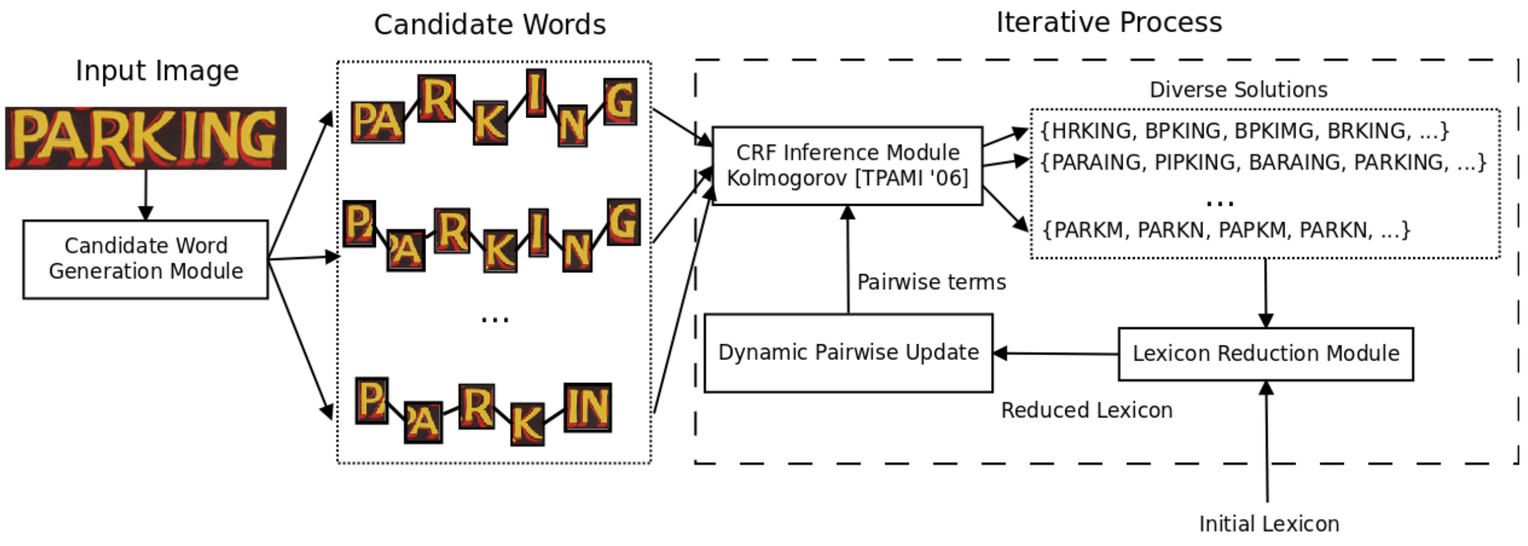
Wang K et al. [ICCV '11], Shi et al. [CVPR '13], Wang T et al. [ICPR '12], Mishra et al. [CVPR '12]

- Potential character locations detected using
  - Binarization
  - Sliding window
- Inference on graphs to recognize text
- Small lexicons used to correct recognition

Drawbacks:

- Difficult to obtain a single set of true character windows in a graph
- Does not perform well on large lexicon settings due to weak pairwise terms

Proposed Method



Lexicon Reduction

Lexicon → / Diverse solutions ↓	STARS	THIS	TAP	...
TARS	**1**	2	**2**	...
TOLS	3	2	3	...
THIS	3	**0**	3	...
Group Edit Distance	1	0	2	...
Rank →	2	1	3	...

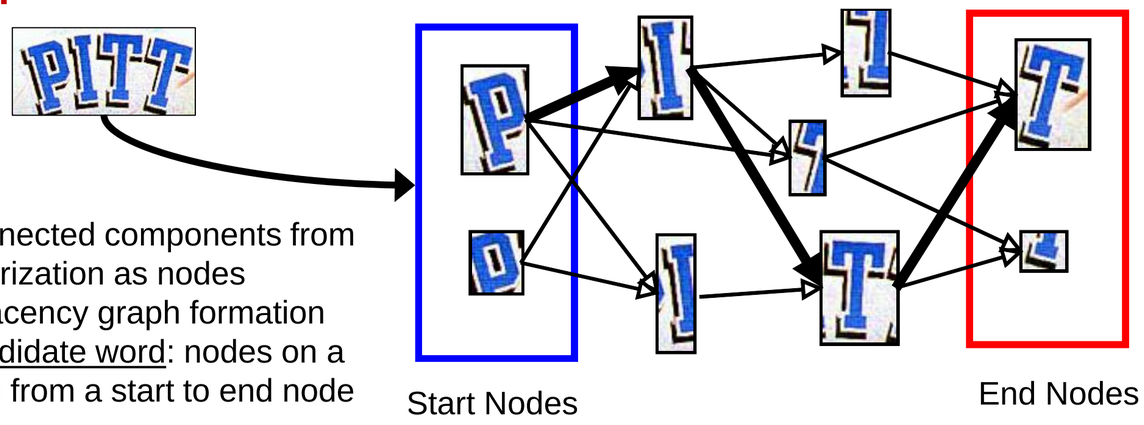
Group edit distance based re-ranking using edit distances between lexicon words and diverse solutions

Lexicon Reduction Process

Given an initial lexicon and diverse solutions,

I. Re-rank using group edit distances
II. Select the top-K words as reduced lexicon

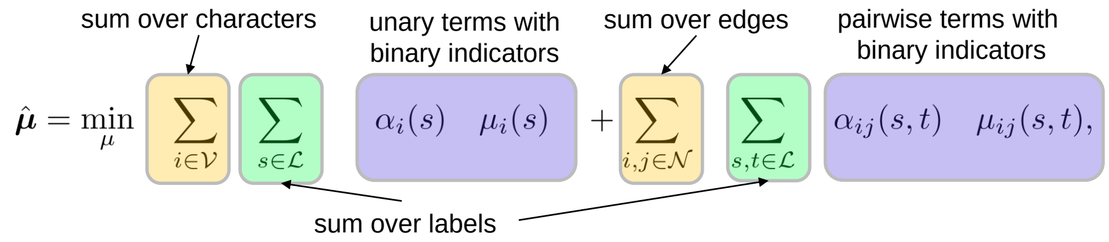
Multiple Candidate Words



- Connected components from binarization as nodes
- Adjacency graph formation
- Candidate word: nodes on a path from a start to end node

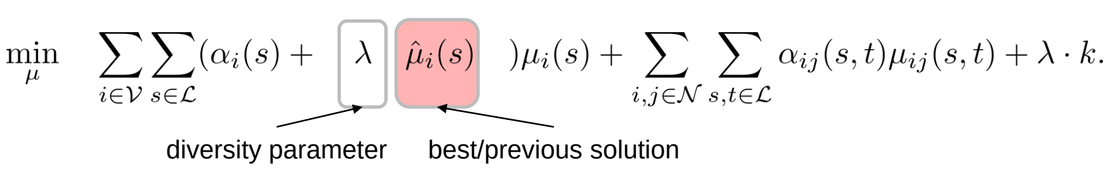
Diverse Solutions

We define a CRF over a candidate word and infer the minimum energy label by optimizing the following,



$$\hat{\boldsymbol{\mu}} = \min_{\mu} \sum_{i \in \mathcal{V}} \sum_{s \in \mathcal{L}} \alpha_i(s) \quad \mu_i(s) + \sum_{i,j \in \mathcal{N}} \sum_{s,t \in \mathcal{L}} \alpha_{ij}(s,t) \quad \mu_{ij}(s,t),$$

We obtain the next solution by adding the constraint $\Delta(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) \geq k$ with at least *k* hamming distance from best solution. Dualizing the constraint,



$$\min_{\mu} \sum_{i \in \mathcal{V}} \sum_{s \in \mathcal{L}} (\alpha_i(s) + \lambda \hat{\mu}_i(s)) \mu_i(s) + \sum_{i,j \in \mathcal{N}} \sum_{s,t \in \mathcal{L}} \alpha_{ij}(s,t) \mu_{ij}(s,t) + \lambda \cdot k.$$

- Second solution obtained by modifying unary terms and inferring again
- Diverse solutions similar to Batra et al. [ECCV '12]
- Process can infer the true label even with a incorrect MAP solution

Recognition

I. Given a word image, find multiple candidate words
II. Iteratively reduce the lexicon to size 10
III. Infer diverse solutions with pairwise terms from reduced lexicon
IV. Find a word in the original lexicon with least group edit distance

K	ICDAR 03		ICDAR 11		ICDAR 13		IIIT 5K-word		SVT	
	M	L	M	L	M	L	M	L	M	L
	Non Diverse									
1	78.9	61.5	69.9	51.1	70.5	51.0	58.0	40.9	66.4	48.3
3	79.1	63.9	69.8	51.4	70.5	51.2	58.3	40.9	66.6	48.9
5	79.2	63.6	69.9	51.2	70.5	51.2	57.8	40.0	66.7	48.8
	Diverse									
1	77.0	62.7	68.2	52.3	69.4	52.6	57.7	38.9	67.2	**51.4**
3	78.3	**66.9**	**70.3**	57.2	70.6	57.1	62.2	43.9	67.2	**51.4**
5	**80.0**	66.5	70.0	**58.1**	**72.1**	**59.0**	**62.9**	**45.3**	**67.3**	**51.4**

Word recognition accuracy comparison over medium and large lexicons in diverse and non-diverse setting for top-K solutions

Method	IIIT 5K-word			ICDAR 03	SVT
	Large	Medium	Small	Small	Small
non-CRF based					
Wang et al. [ICCV 2011]	-	-	-	76.0	57.0
Bissacco et al. [ICCV 2013]	-	-	-	82.8	**90.3**
Alsharif et al. [arXiv 2013]	-	-	-	**93.1**	74.3
Goel et al. [ICDAR 2013]	-	-	-	89.6	77.2
Rodriguez et al. [BMVC 2013]	-	57.4	**76.1**	-	-
CRF based					
Shi et al. [CVPR 2013]	-	-	-	87.4	73.5
Novikova et al. [ECCV 2012]	-	-	-	82.8	72.9
Mishra et al. [CVPR 2012]	-	-	-	81.7	73.2
Mishra et al. [BMVC 2012]	28.0	55.5	68.2	80.2	73.5
Our Method	**45.3**	**62.9**	71.6	85.5	76.4

Word recognition accuracy comparison between various CRF and non-CRF methods

Retrieval

Pre-processing stage

I. Reduce lexicons for each word image to size 4
II. Compute average edit distance(AED) by averaging the edit distances of all word pairs in the lexicon
III. If AED ≥ θ, then reduce lexicon to size 1 else keep lexicon of size 4

Retrieval stage

I. Retrieve images with the query word in their reduced lexicon
II. Rank the images based on lexicon sizes and query word position

Method	IIIT 5K-word			ICDAR 03
	Large	Medium	Small	Small
Without diversity				
Full Reduction	27.5	51.9	65.0	**81.7**
Partial Reduction	35.1	35.6	60.7	76.9
With diversity				
Full Reduction	23.1	52.0	65.0	78.9
Partial Reduction	**42.1**	**59.0**	**66.5**	79.5

Top-1 precision results on various datasets

Query	Retrieved Image	Reduced Lexicon: diversity + partial red.	Reduced Lexicon: diversity + full red.
BRADY		MY, BRADY, ANY, A	MY
SPACE		HOT, SPACE, LACEY, SALE	HOT
HAHN		BUENA, HANDA, HAHN, PIPE	BUENA
DAILY		PEARL, MOUNTS, DAILY, NIKE	PEARL
TIMES		TIME, TIMES, WINE, MED	TIME
THREE		THE, THREE, THERE, USED	THE

Correct retrieval cases using partial reduction and diverse solutions

Center for Visual Information Technology
International Institute of Information Technology, Hyderabad, India
http://cvit.iiit.ac.in/projects/SceneTextUnderstanding

* LEAR team, Inria Grenoble Rhône-Alpes, Laboratoire Jean Kuntzmann, CNRS, Univ. Grenoble Alpes, France.