WordFences: Text Localization and Recognition

ICIP 2017

Andrei Polzounov (Universitat Politecnica de Catalunya, Barcelona, Spain),

Artsiom Ablavatski (A*STAR Institute for Infocomm Research, Singapore),

Dr Sergio Escalera (Universitat de Barcelona, Barcelona, Spain),

Dr Shijian Lu (A*STAR Institute for Infocomm Research, Singapore),

Dr Jianfei Cai (Nanyang Technological University, Singapore)

Sponsors

- Institute for Infocomm Research (I^2R), at Singapore's Agency for Science, Technology and Research (A*STAR)
- CERCA Program, Government of Catalonia

Problem Description

- Text detection and recognition in natural scene imagery.
- Good test case problem for AI research and for uses in industry: mapping business from StreetView, translating menus or billboards, *etc.*

Motivation

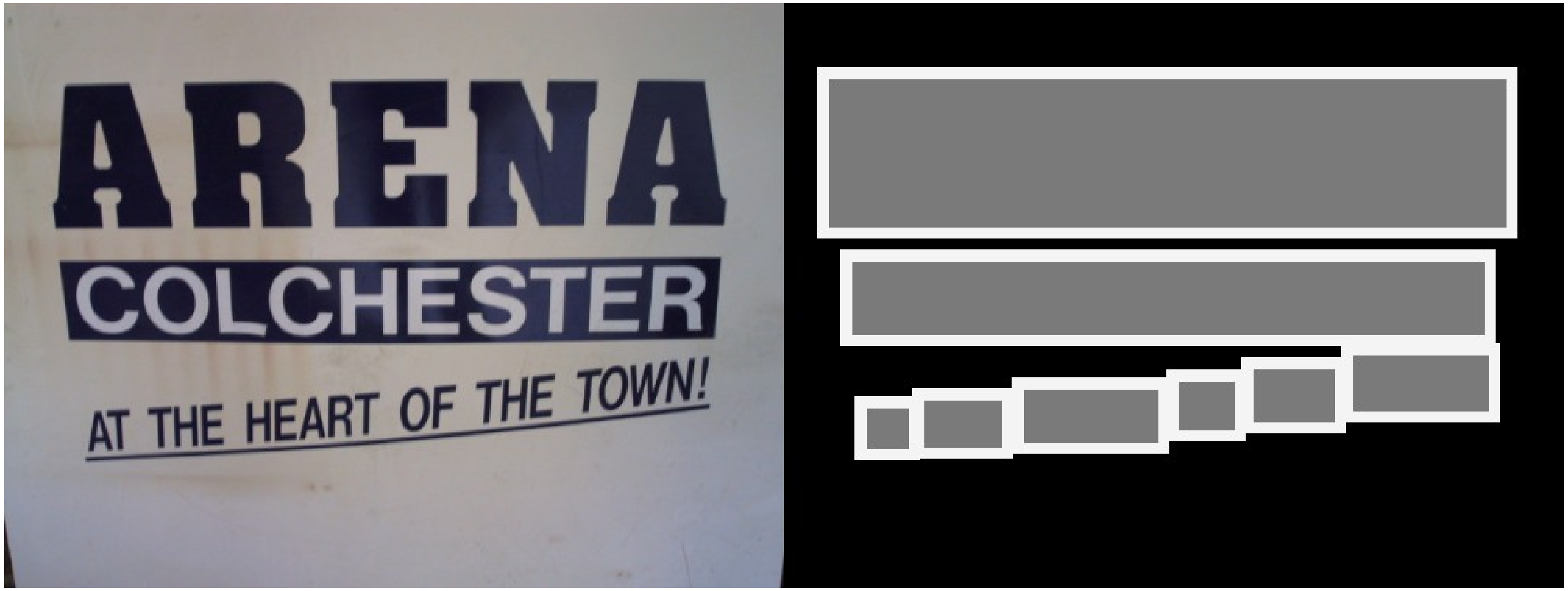
- OCR can be used on scanned text.
- Natural images have a ton of variety in fonts, scales, kerning and features.

Proposed Solution

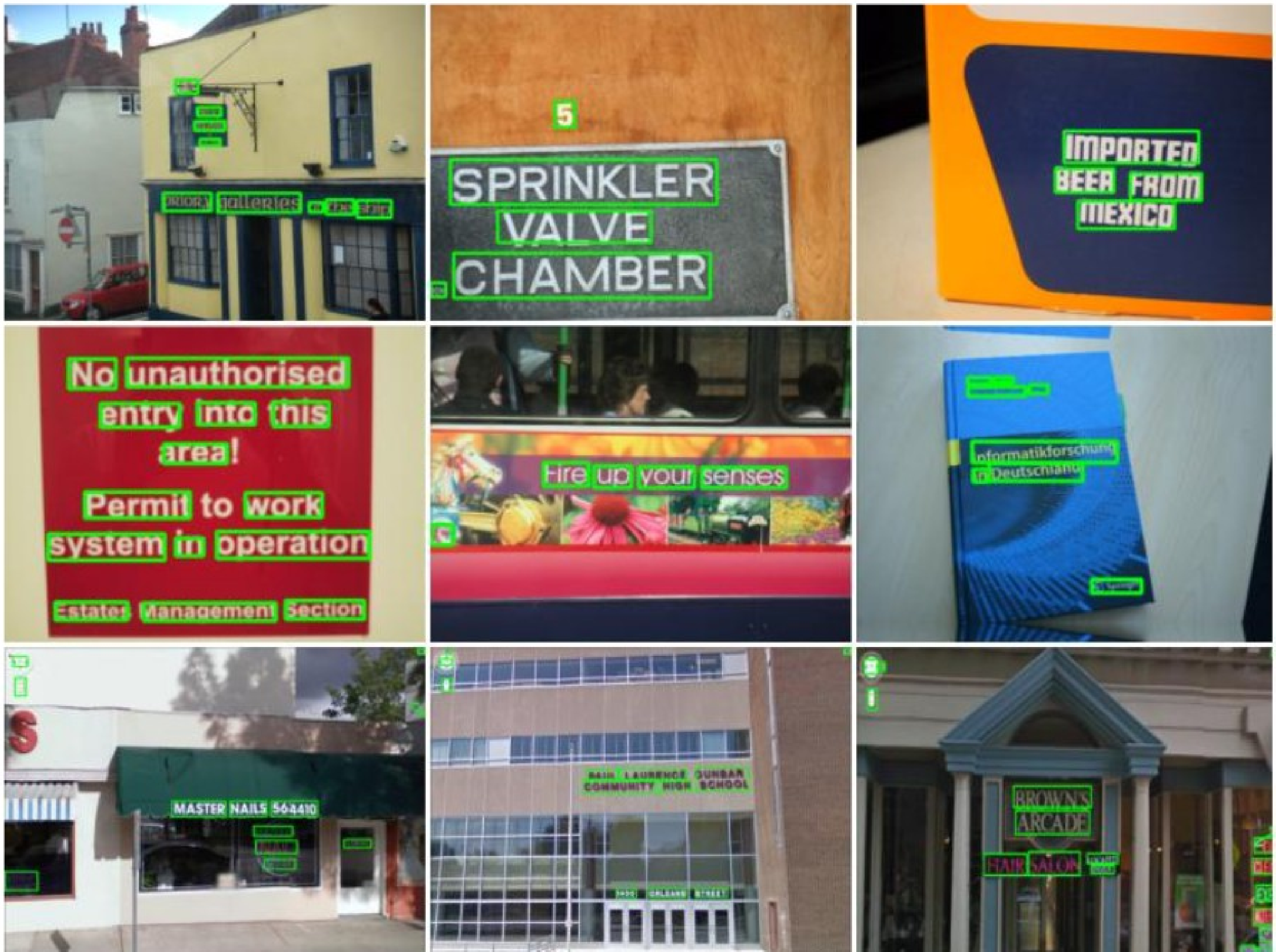
- 2-stage deep learning network
  - Find locations/ROIs (text localization) with CNN.
  - Detect characters (end-to-end text recognition) with RNN.
- 1st stage is the more difficult one. It is related to object recognition and semantic segmentation problems in Computer Vision.

Contributions

- Two main contributions:
  - Use a word separator – a „WordFence“ for deliniating individual words.
  - Using a novel weighted pixelwise softmax function for training the semantic segmentation.

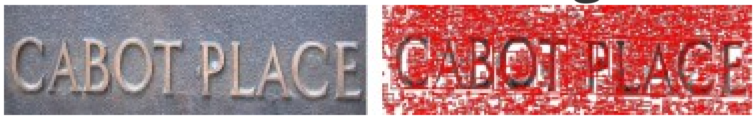


Sample Detections



Related Work (CV)

- Maximally Stable Extremal Regions by Huang *et al.*(2014) – works by first using an MSER transform and then training a CNN.

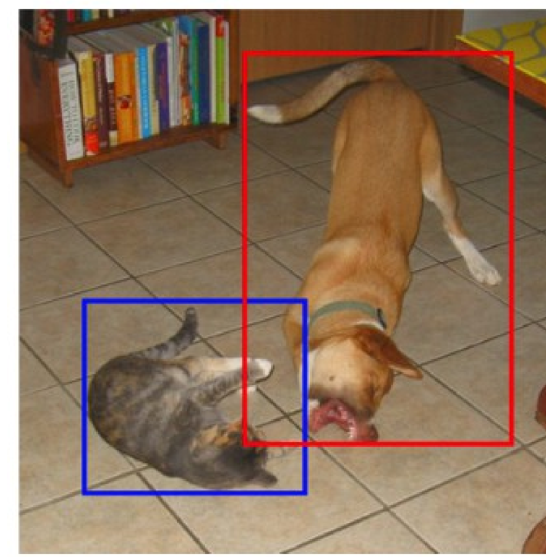


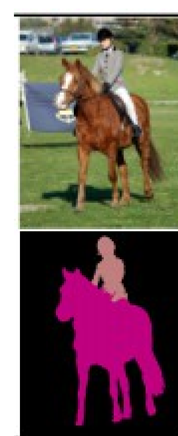
- Stroke Width Transform by Epshtein et al. (2010) – an edge detection method, that relies on the fact that a given text font should have similar width/thickness for each stroke in a character.

- Edge Boxes by Zitnick and Dollár (2014) – simple object score based on the number of edges within a given sliding window. Sparse and fast to evaluate, but results could be better.

Related Deep Learning

- Single shot ROI detection: YOLO (2015) by Redmon *et al.* and SSD: Multibox (2015) by *Liu et al.* - both of these work by splitting image into grid and calculating probabilities
- Fully convolutional networks (2015) by Long and Shelhamer – replace fully connected layers with deconv.
- DeepLab (2015) Liang-Chieh *et al.* - and ResNet101 (2016) by He *et al.* are the bases for this work.



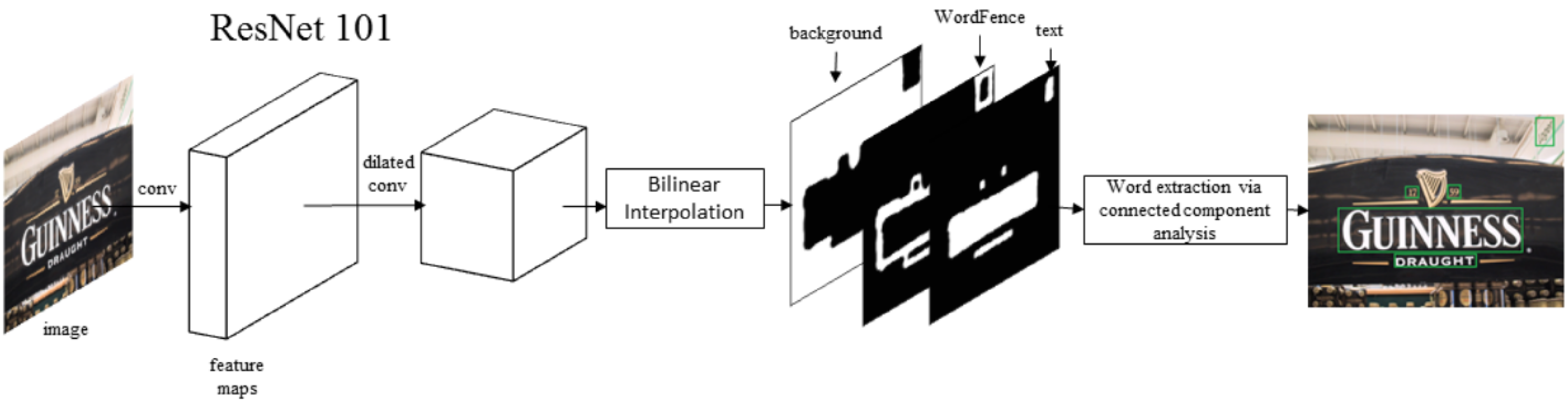


Related Text Detection

- SynthText (2016) by Gupta, Vedaldi, and Zisserman. Synthetic dataset for text recognition and one-shot box approach to learning.

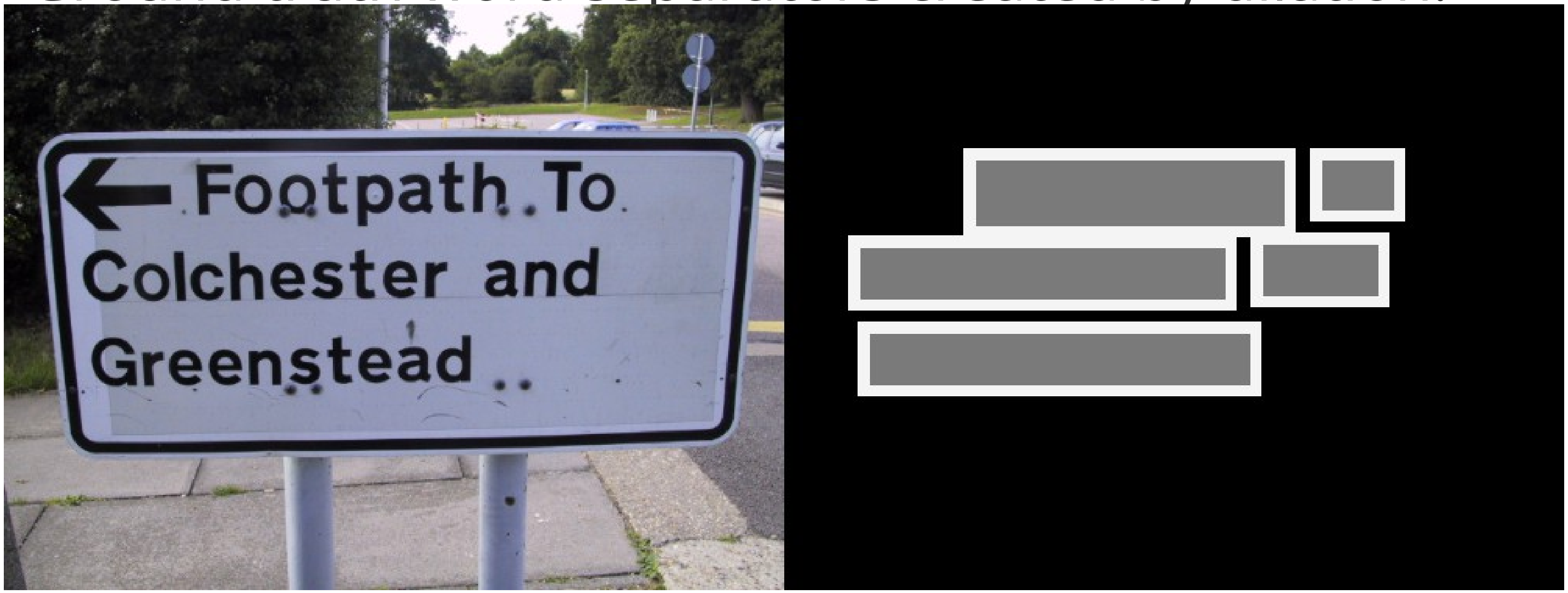
- Jaderberg *et al.* (2014) – used CNN to generate high number of proposals and filter with Random Forest (HoG).
- He *et al.* (2016) – cascaded CNN with false positive rejection to detect textlines (no split).

Model Overview



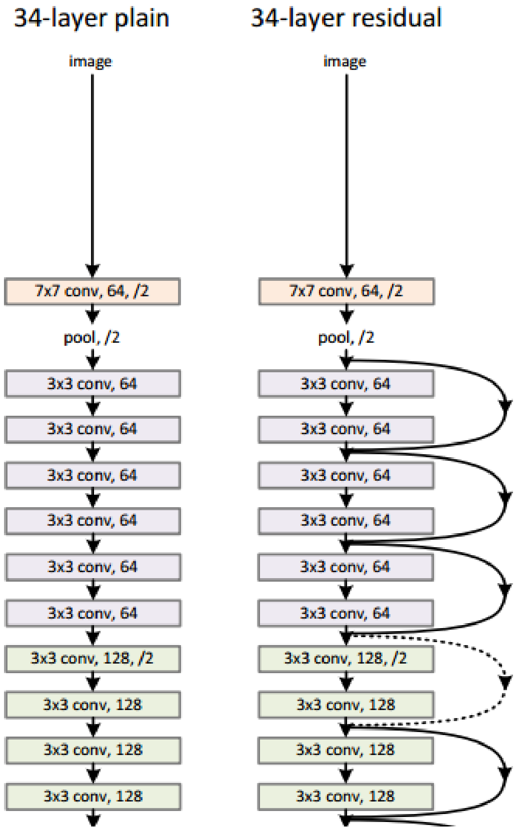
Word Localization as Semantic Segmentation

- Semantic segmentation is a well known problem.
- Able to handle different scales using wide fields of view and multi-scale inference.
- Ground truth word separators created by dilation.



ResNet of Exponential Receptive Fields

- ResNets help to train huge networks without a vanishing gradient.
- Receptive fields can be enlarged using convolutional dilations in the ConvNets.
- Deep network + exponential receptive fields = effective multi-scale detection of different sized text.



Weighted Pixelwise Softmax Loss

- Background pixels are the majority. More emphasis for text and WordFence pixels is needed.

Algorithm 1 Pixelwise Weighted Softmax Loss

Require: Predicates after fusion $\mathbf{Pr}$, ground truth labels $\mathbf{L}$

1: $probs \leftarrow Softmax(\mathbf{Pr})$ ▷ pixel probabilities

2: $m \leftarrow NumberOfUniqueLabels(\mathbf{L})$

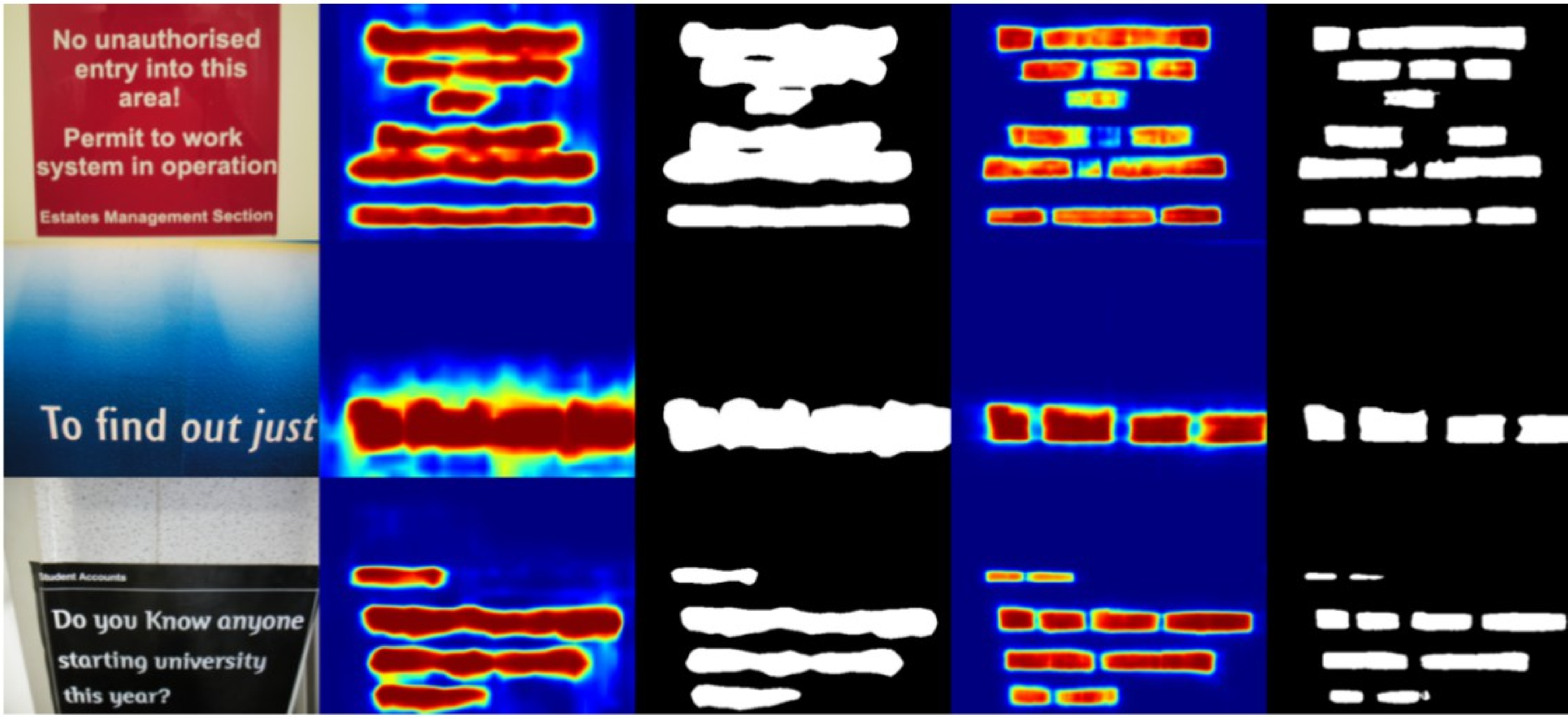
3: $n_1, n_2, \ldots, n_m \leftarrow CountsOfUniqueLabels(\mathbf{L})$ ▷ get counts of each label on a ground truth image

4: $loss \leftarrow -\sum \frac{1}{n_{gt}} \log(probs_{gt})$ ▷ weighted loss calculation

5: $Backpropagate(loss, \frac{1}{n_1}, \frac{1}{n_2}, \ldots, \frac{1}{n_m})$ ▷ loss backpropagation with normalization factors

- Algorithm allows us to rebalance per image weights on the fly.

WordFence vs No-WordFence



- WordFences act as penalization for merged words.

Text Datasets

- ICDAR 2011 and ICDAR 2013 – International Conference on Document Analysis and Recognition
- COCO-Text – subset of the popular MS-COCO dataset for object recognition (21 classes)
- SVT – Google Street View data
- SynthText – synthetic text mixed with scene images from Gupta *et al.*

Localization Results

Model	PASCAL VOC IoU = 0.5								
	ICDAR11			ICDAR13			SVT		
	P	R	F	P	R	F	P	R	F
Tian *et al.* [32]	0.89	0.79	0.84	0.93	0.83	**0.88**	-	-	-
Gupta *et al.* [5]	0.78	0.63	70.0	0.78	0.63	0.70	0.47	0.45	0.46
Jaderberg *et al.* [11]*	0.89	0.68	77.4	0.89	0.68	0.77	0.59	0.49	0.54
Gupta *et al.* [5]*	**0.94**	0.77	**0.85**	**0.94**	0.76	0.84	**0.65**	0.60	**0.62**
WDN (ours)	0.64	**0.92**	0.75	0.65	**0.92**	0.76	0.47	**0.63**	0.54

- Methods marked with * use multi-stage false-positive detectors.

Localization Results

Model	PASCAL VOC IoU = 0.5								
	ICDAR11			ICDAR13			SVT		
	P	R	F	P	R	F	P	R	F
Tian *et al.* [32]	0.89	0.79	0.84	0.93	0.83	**0.88**	-	-	-
Gupta *et al.* [5]	0.78	0.63	70.0	0.78	0.63	0.70	0.47	0.45	0.46
Jaderberg *et al.* [11]*	0.89	0.68	77.4	0.89	0.68	0.77	0.59	0.49	0.54
Gupta *et al.* [5]*	**0.94**	0.77	**0.85**	**0.94**	0.76	0.84	**0.65**	0.60	**0.62**
WDN (ours)	0.64	**0.92**	0.75	0.65	**0.92**	0.76	0.47	**0.63**	0.54

- Methods marked with * use multi-stage false-positive detectors.

Recognition Results

Model	Year	IC11	IC13
Neumann *et al.* [20]	2013	0.45	-
Jaderberg *et al.* [11]	2015	0.69	0.76
Gupta *et al.* [5]	2015	**0.84**	0.85
WDN	2016	**0.84**	**0.86**

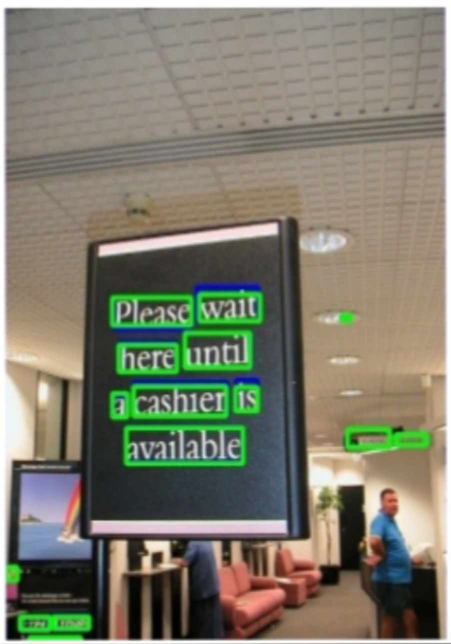
- The recognition stage uses proposals given from the detection network.
- Recognition stage is based on the CRNN network by Shi *et al.* (2016).

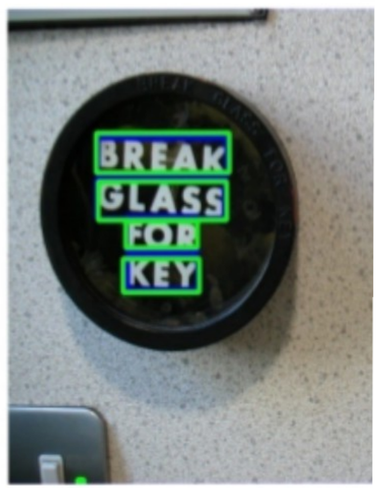
Localization Results

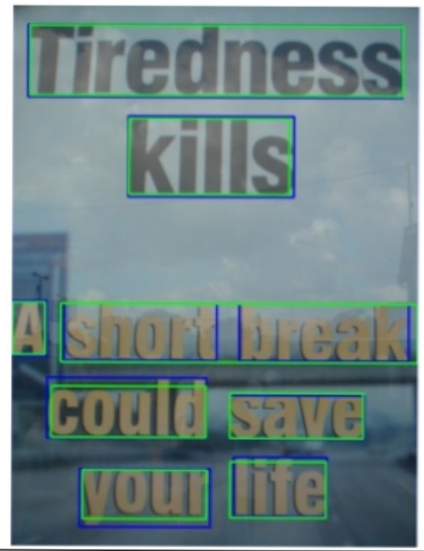
Model	PASCAL VOC IoU = 0.5								
	ICDAR11			ICDAR13			SVT		
	P	R	F	P	R	F	P	R	F
Tian *et al.* [32]	0.89	0.79	0.84	0.93	0.83	**0.88**	-	-	-
Gupta *et al.* [5]	0.78	0.63	70.0	0.78	0.63	0.70	0.47	0.45	0.46
Jaderberg *et al.* [11]*	0.89	0.68	77.4	0.89	0.68	0.77	0.59	0.49	0.54
Gupta *et al.* [5]*	**0.94**	0.77	**0.85**	**0.94**	0.76	0.84	**0.65**	0.60	**0.62**
WDN (ours)	0.64	**0.92**	0.75	0.65	**0.92**	0.76	0.47	**0.63**	0.54

- Methods marked with * use multi-stage false-positive detectors.

ICDAR 2011

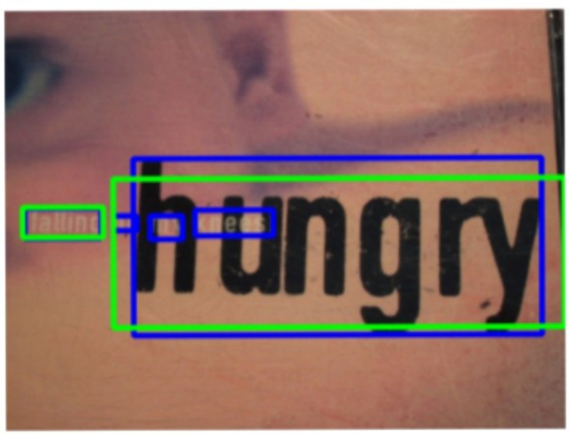


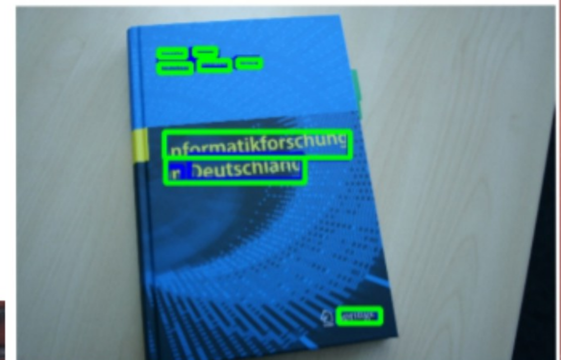


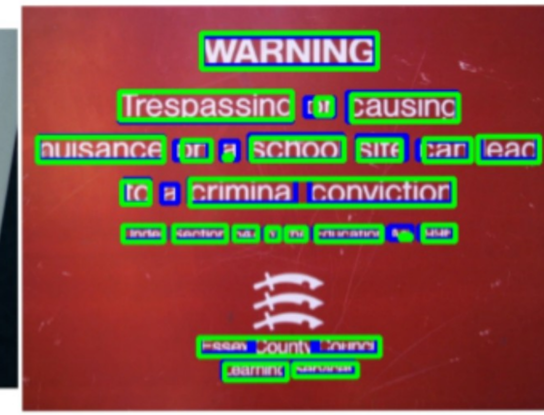


P=0.875, R=0.778, F=0.824	P=0.625, R=1.000, F=0.769
"short break" not split correctly	Some false positives

ICDAR 2013







P=1.000, R=0.889, F=0.941	P=0.833, R=0.714, F=0.769
	Some overlaps are <0.5 IoU

P=1.000, R=0.400, F=0.571	P=0.300, R=0.750, F=0.429
Challenging case - word is within another word	Too much word splitting, due to tricky font and spacing

Problems Encountered

- Going from segmentations to bounding boxes.
- Noisy detections and many false positives:
  - Many small detected regions.
  - Camera artifacts such as glare.
  - Need for balancing precision vs recall.

Conclusion

- Text recognition as semantic segmentation
- WordFences as penalization
- SOTA recall on detection, which provides high quality samples to the recognition stage (which in itself is able to throw away false positives).
- SOTA F-scores on recognition
- Paper submitted to ICIP 2017 – International Conference on Image Processing

Future Work

- WDN relies on visual information to split words.
- Humans also use word semantics and memory.
  - “Raeding wrods with jubmled letetrs“.
- A smarter system would be able to read text directly from images without a midpoint CV representation.

Questions?

UPC
UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

UNIVERSITAT
ROVIRA I VIRGILI

UB
Universitat de Barcelona

A*STAR
Institute for
Infocomm Research