

Universitat Politècnica de Catalunya
Universitat Rovira i Virgili
Universitat de Barcelona

Facultat d'Informàtica de Barcelona
Campus Nord Building B6
C/Jordi Girona, 1-3,
Barcelona, Spain
08034



UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH



UNIVERSITAT
ROVIRA I VIRGILI



Universitat de Barcelona

Computer Science Master Thesis

WordFences:
Text Localization and Recognition

Andrei Polzounov

January 2017

Academic Supervisor:

Dr. Sergio Escalera

Universitat de Barcelona, Barcelona, Spain

Industry Supervisor:

Dr. Shijian Lu

Institute for Infocomm Research, Singapore



A*STAR, Institute for Infocomm Research (I²R)
1 Fusionopolis Way,
Singapore 138632

This dissertation originated in cooperation with the Institute for Infocomm Research (I²R), at Singapore's Agency for Science, Technology and Research (A*STAR).

First of all I would like to thank Dr. Shijian Lu at I²R for giving me the opportunity to carry out state of the art research in this field and for hosting me at I²R in Singapore for the duration of my research.

谢谢

Special thanks to Mr. Artsiom Ablavatski for his help at I²R.

Furthermore I would like to thank the Bothans for their sacrifice in obtaining the Death Star plans.

Abstract

In recent years, text recognition has achieved remarkable success in recognizing scanned document text. However, word recognition in natural images is still an open problem, which generally requires time consuming post-processing steps. We present a novel architecture for individual word detection in scene images based on semantic segmentation. Our contributions are twofold: the concept of WordFence, which detects border areas surrounding each individual word and a unique pixelwise weighted softmax loss function which penalizes background and emphasizes small text regions. WordFence ensures that each word is detected individually, and the new loss function provides a strong training signal to both text and word border localization. The proposed technique avoids intensive post-processing by combining semantic word segmentation with a voting scheme for merging segmentations of multiple scales, producing an end-to-end word detection system. We achieve superior localization recall on common benchmark datasets - 92% recall on ICDAR11 and ICDAR13 and 63% recall on SVT. Furthermore, end-to-end word recognition achieves state-of-the-art 86% F-Score on ICDAR13.

Keywords

machine learning, artificial intelligence, object detection, text detection, text recognition

Thesis Domain (Erasmus subject area code)

11.4 Artificial Intelligence

Contents

1. Introduction	1
1.1. Motivation	1
1.2. Background	1
1.3. Contribution	3
1.4. Outline	3
2. Related Work	4
3. WordFence Detection Network Model	6
3.1. Overview	6
3.2. Word Localization as Semantic Segmentation	6
3.3. ResNet of Exponential Receptive Fields	7
3.4. Weighted Softmax Loss Function	7
4. Evaluation and Experiments	10
4.1. Datasets	10
4.2. Word Detection Experiments	10
4.3. Timings	11
4.4. End-to-end Word Detection and Recognition	12
5. Conclusion	13
5.1. Summary	13
5.2. Dissemination	13
5.3. Problems Encountered	13
5.3.1. Segmentations to bounding boxes	14
5.3.2. Noisy detections and false positives	14
5.4. Future Work	14
List of Acronyms	15
Appendices	19
A. Text Detection	19
A.1. ICDAR 2011	20
A.2. ICDAR 2013	22
A.3. SVT	24

List of Figures

1.1. Word detection bounding box results on ICDAR2011 (top), ICDAR2013 (middle) and SVT (bottom) datasets. Bounding boxes are the output of the proposed method.	2
3.1. WordFence detection network architecture.	6
3.2. Word localization as semantic image segmentation. Individual words are trained to be split up using purely visual information. Left and middle-left: confidence levels as heatmaps for text and WordFence area. Mid-right: segmentation results. Far-right: word bounding box obtained from segmentation by connected component analysis.	8
3.3. Comparisons of segmentation with and without WordFence. The first column from the left shows the original images. The second and third show the text position belief map and the resulting segmentation, respectively. Last two columns show the belief map and the segmentation from our method. Localizing words without WordFence has a tendency of individual words bleeding over into each other, which causes difficulties to posterior word recognition.	9

List of Tables

4.1. Description of the text recognition datasets	10
4.2. Comparisons with other methods of word detection. Precision, Recall and F-Score are reported. Recall maximization was necessary for obtaining good word detection results. Methods marked with * use a multistage false-positive filtering process to increase precision, the code was not published thus the results are not directly comparable with ours.	11
4.3. Comparison of word detection time (in seconds)	12
4.4. Evaluation of end-to-end word recognition on ICDAR 2011 and 2013 datasets. F-score is reported.	12
1. ICDAR11 Results Part 1	20
2. ICDAR11 Results Part 2	21
3. ICDAR13 Results Part 1	22
4. ICDAR13 Results Part 2	23
5. SVT Results: The SVT dataset ground truth is not well-labeled and the experimental results on this dataset do not necessarily correspond to a good localization network. However, a visual, qualitative inspection of the results shows that WDN detects and splits the majority of words present. It is also interesting to note that the Google Maps compass is often detected as a word - showing the same trend of detecting pictorials as in Table2	24

1. Introduction

It's the repetition of affirmations that leads to belief. And once that belief becomes a deep conviction, things begin to happen.

Muhammad Ali

Machine reading of text in images has long attracted interest. Text recognition in natural images has presented great challenges until now. Recent developments in artificial intelligence, machine learning, computer vision and natural language processing have all contributed to advances in text recognition in natural images. Text recognition provides exciting potential both as a test case problem for artificial intelligence and for its many potential applications in industry: mapping new businesses from Google Street View, translating menus or billboards from foreign languages, *etc.*

1.1. Motivation

Detection and recognition of text in natural images has long been an outstanding challenge in the computer vision and machine learning communities. Text recognition in the wild can provide context and semantic information for scene understanding, object classification and action recognition in images or video. The task has attracted interest of many researchers [5, 12, 11, 32, 31, 44, 7]. Due to the difficulty of text detection in natural images, even state-of-the-art systems struggle with word localization because of the staggering variety of text sizes and fonts, potentially poor image quality, low contrast, image distortions, or presence of patterns visually similar to text such as signs, icons or textures. Many works in text detection employ knowledge-based algorithms and heuristics in order to tackle these challenges. Some of the most common techniques include: text line extraction [7, 32], character candidate detection [31, 9] and using secondary classifiers to remove false positive detections [11].

1.2. Background

Recent successes in computer vision are centered on deep convolutional neural networks (CNNs). Since the seminal 2012 paper by Krizhevsky *et al.* [13], that won the ImageNet competition [25], deep learning and convolutional neural networks in particular have been the focus of many researchers. Some of the problems being solved are: object-classification in natural images [13, 29], pixelwise semantic segmentation [18, 15, 21, 43,

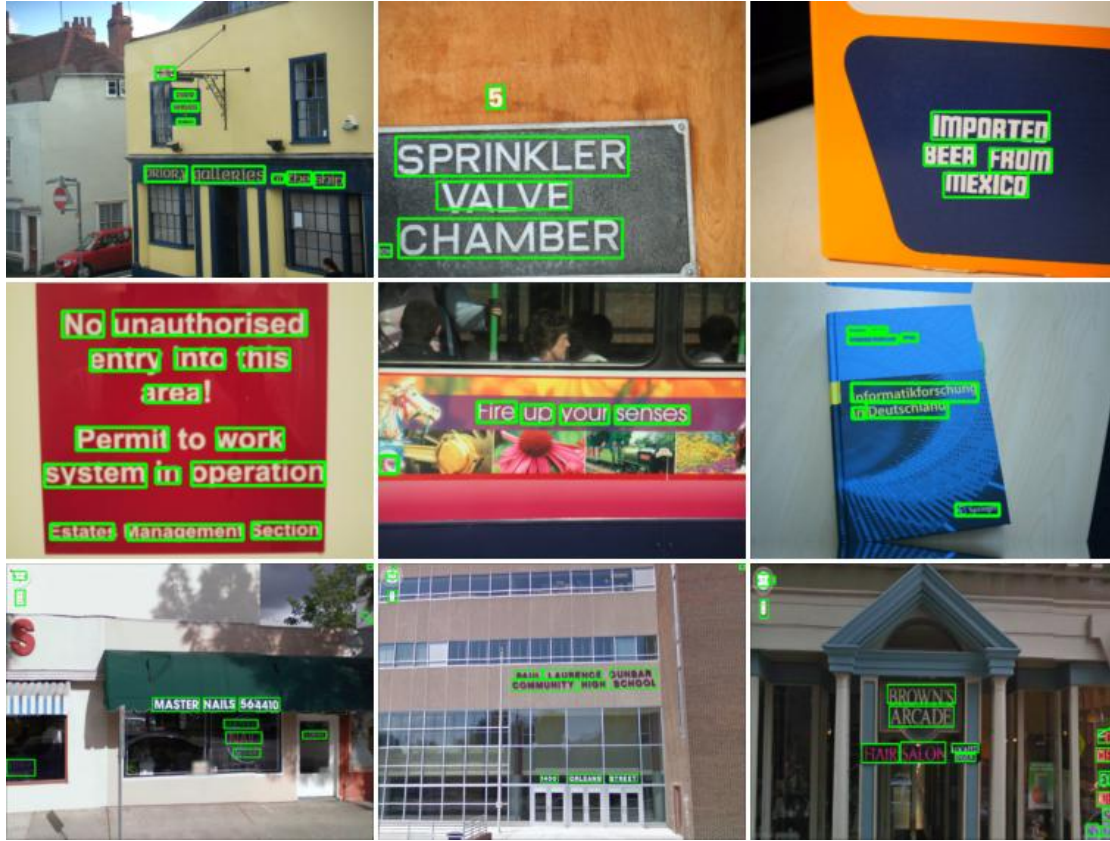


Figure 1.1.: Word detection bounding box results on ICDAR2011 (top), ICDAR2013 (middle) and SVT (bottom) datasets. Bounding boxes are the output of the proposed method.

41], human pose estimation [39, 33], bounding box detection [23, 17, 45, 26], and text detection in scene images [11, 5, 42, 44, 38, 8].

A major limitation of CNNs is that networks have trouble taking different scales of images into account when analyzing objects of different sizes. Modern CNNs use max-pooling layers to reduce resolution and search space for training - this operation reduces resolution and loses spatial information between different features. Yu and Koltun [41] have argued that max-pooling operations do not maintain sufficient global scale information and thus proposed dilated convolutions to increase the effective receptive field of convolutional operations without a losing resolution. Other works tackled the scale problem with methods such as fully convolutional networks (FCNs) [18] or with atrous convolutions [26, 15].

Another challenge that has been addressed by CNNs is semantic segmentation - problems where each pixel in the image must be matched to a specific label. Semantic segmentation has recently been enhanced by dilated convolutions [41], FCNs [18] and

probabilistic graphical models, such as conditional random fields [15].

1.3. Contribution

In this work, we treat the task of word detection as a semantic segmentation problem of three separate classes: words, background and WordFences (Section 3.2). After detecting an area of interest we compute bounding boxes for each word-proposal in the image. Most current state-of-the-art region of interest (ROI) detectors like Faster-RCNN (F-RCNN) [24] use a variation of the following steps: propose bounding boxes, resample pixels of the ROI and then apply a second classifier to filter and improve proposals. In contrast with F-RCNN, our high quality segmentation results allow us to extract accurate bounding box proposals directly from the segmentation. The segmentation maps are obtained by inference at different image scales and combining the results with an efficient voting mechanism. Merging the results from different scales further helps to eliminate duplicate proposals for the same word and to remove most false positive detections.

The WordFence detection network (WDN) is designed to take advantage of deep residual connections and whole-image receptive fields. By treating the text localization problem as a semantic segmentation and training with a self-penalizing loss function which recalculates class weights on the fly (see Section 3.4) we achieve good detection results with superior recall. The new model achieves state-of-the-art text detection performance on benchmark datasets (see Section 4.1) while avoiding the need of knowledge-based approaches such as text lines. End-to-end word recognition achieves state-of-the-art F-Score of 86% on ICDAR13 (see Section 4.4). Sample word detection results are presented in Fig. 1.1 and in Appendix A.

1.4. Outline

This thesis is separated into 5 chapters.

Chapter 2 describes related work in object (and text) recognition, detection and classification. Recent advances in both object and text detection are reviewed and presented.

Chapter 3 provides a summary of the machine learning components necessary for building a text detection and recognition pipeline. Concepts of semantic segmentation for text area recognition and ROI extraction are introduced and provide reference to the rest of the thesis.

Chapter 4 contains the evaluation results based on well known benchmark datasets. An overview of the datasets and comparisons with related networks are provided.

Chapter 5 summarizes the thesis and suggests future research directions.

2. Related Work

Traditionally text recognition has focused on documents, and several optical character recognition (OCR) techniques have been developed for this task. Recently, text detection in scene imagery has come to the forefront. Generally, text recognition works by first providing a “candidate bounding box” - or a proposal for a single word or a word-line. The word proposal is then cropped out of the natural image and fed to a word recognition network which then matches words against an internal dictionary.

In the aforementioned scenario, text localization is considered to be the key task, since a well-cropped proposal can be fed to many well known word recognition systems [10]. Before CNNs, popular methods for text localization utilized classical computer vision techniques such as sliding windows with hand-crafted feature descriptors. More recent works have utilized CNN features, some prominent methods include: Maximally Stable Extremal Regions [9], Stroke Width Transform [2], and EdgeBoxes [45], and others. These methods feature a combination of character recognition CNN and a sliding window algorithm. Amongst the feature driven techniques, the most impressive results, and the current state-of-the-art, were achieved by Tian *et al.* [32] by using a combination of a CNN for individual character detection and graph optimization techniques. However, all of these approaches have a general limitation of feature driven engineering - there are simply too many edge cases to account for. The detectors generate a large amount of non-text false positives (pictorials, signs, bricks and other textures may appear to be similar to text to the neural network), raising the problem of creating additional filtering techniques and work-arounds. Often, a number of post-processing steps is needed to reach a good performance.

With the prominence of deep learning, CNN based regression of candidate bounding boxes have started being utilized for filtering false positive candidates. Bounding box detection has been proposed in the context of object detection by works such as YOLO [23], F-RCNN [24] and SSD: Single Shot MultiBox Detector [17]. Advances in semantic segmentation [42, 7] have allowed dense prediction to provide input to bounding box generation algorithms. Building on successful implementations of CNNs for semantic segmentation using FCNs for dense prediction [18, 43], several researchers have introduced object localization with FCNs [14, 21]. For example, F-RCNN [24] utilizes an FCN for accurate bounding box proposal generation. However, Gidaris and Komodakis [3] found that regressing bounding boxes directly from neural network parameters constitutes a complicated learning task that may not provide accurate bounding boxes. Accuracy in bounding boxes for word proposals is doubly important in text localization and recognition. Word proposals with characters cut off or proposals with overlap over other words would significantly inhibit the word recognition stage of an end-to-end learning system (see Section 4.4 for our end-to-end testing results).

With the growing popularity of neural networks and end-to-end pipelines, the computer vision community has shifted its focus on to the general problem of object detection [24, 26, 23, 17]. Most object proposal generators follow the blueprint of end-to-end supervised training on a big amount of labeled data for direct prediction of object bounding boxes in the image with a single step inference. Although detectors show impressive results and surpass previous work in object detection, these methods may not produce accurate localization results. Usually, post processing steps are needed to further refine object localization.

Reinforcement learning is another interesting area of research to text localization. Caicedo and Lazebnik [1] trained a network for iterative object localization based on reinforcement learning principles. An improved version of their approach was developed by Lu *et al.* [19]. However, due to high data variability and complicated reward functions these works do not outperform state-of-the-art object detection results, but this is an interesting direction for further research and this field could potentially prove fruitful with larger datasets.

Several text detection works have been inspired by recent object detection approaches. Early work by Zhang *et al.* [42] used a semantic segmentation model to extract text proposals and refine them by applying hand-crafted heuristics. He *et al.* [7] improved on previous approaches by introducing an additional CNN for refinement, by building a cascade of networks. Two networks were trained separately, and were executed in series for evaluation. Gupta *et al.* [5] adapted YOLO’s approach [23] for text detection and introduced SynthText - a new synthetic dataset (see Section 4.1) proving that a model trained on synthetic data could generalize to real world scenes. Although the model was trained on large amount of data, three post processing steps were applied in order to achieve a good performance. Analogously, F-RCNN [24] was adapted for text recognition by Zhong *et al.* [44] and Tian *et al.* [32]. The former integrated the F-RCNN framework into a more powerful model and added several improvements into the filtering stage. However, the resulting three hundred proposals per image (on average) were then filtered with a time consuming process. Tian *et al.* [32] fused F-RCNN with a recurrent neural network (RNN), allowing the RNN to consider the proposals as a sequence and unite them into text lines. Although He *et al.* [7] also relied on semantic segmentation between text and non-text regions, they did not have a word border area which often caused the resulting segmentations to bleed together (see Fig 3.3). Our approach allows for clean segmented regions without the need for textlines.

Our proposed architecture is inspired by previously mentioned works, but it allows to perform bounding box detection in a single step. Instead of producing a highly non-linear bounding box coordinate prediction as in YOLO [23] and F-RCNN [24], our network takes advantage of semantic segmentation to produce a dense pixel labeling map. Afterwards, word proposals are extracted from the given heat map in a linear time. An overview of the proposed system is shown in Fig. 3.1.

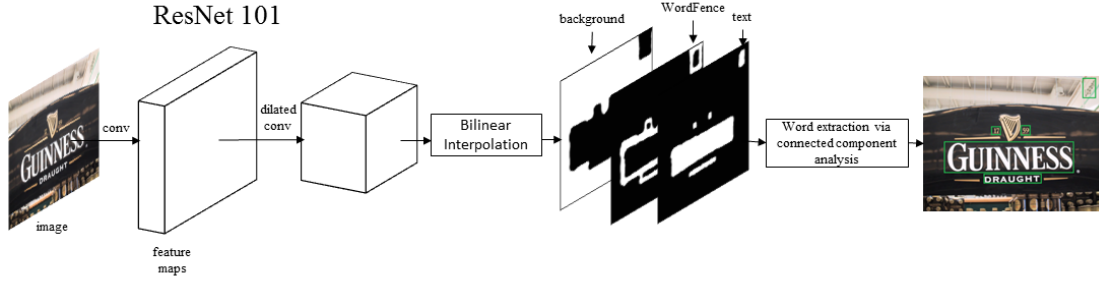


Figure 3.1.: WordFence detection network architecture.

3. WordFence Detection Network Model

3.1. Overview

Inspired by the success of deep ConvNets with residual connections (ResNets), such as the one for semantic segmentation by Chen *et al.* [15], WDN takes advantage of neural network research of the past few years to produce highly accurate detection results. The overall architecture is illustrated in Fig. 3.1. The network includes a ResNet-101 (101 layer residual network introduced by He *et al.* [6]), followed by a number of dilated convolutions [41] that add full-image context to the final classification, before finally performing a bilinear interpolation on the resulting belief map. After the interpolation, connected components are extracted. Each component represents a standalone word on the image which is further processed in the recognition step. Bounding boxes are then extracted from these connected components (see Figure 3.1).

The process of word localization as semantic segmentation is further demonstrated in Figure 3.2.

3.2. Word Localization as Semantic Segmentation

Object segmentation, has recently been considerably improved with the introduction of the deconvolutional layer [18], dilated convolutions (increasing effective receptive field) [41], etc. Significant improvements of resulting metrics and accurate segmentation of objects in difficult scenes have been demonstrated. Several published works [42, 7] have adapted object segmentation for text localization. Semantic segmentation for text localization, despite showing promising results, has had trouble distinguishing individual

words from segmented images. Generally, post processing methods and heuristics were applied to refine word localization results, or the task was not addressed at all as in the case of textline approaches.

We have tackled the challenge of separating each unique word in our segmentation results by our unique method of training the semantic segmentation network. Our training method provides the following new techniques:

- Penalization by training to detect a pixel border between individual words in the image segmentation.
- A unique pixelwise weighted loss function to weigh the words, background and word-separation border regions with an equal weight (see Algorithm 1).

3.3. ResNet of Exponential Receptive Fields

ResNets have achieved great success in recent computer vision tasks [6, 15], surpassing human accuracy. Their depth and structure allow ResNets to train very deep neural networks without a vanishing gradient. Veit *et al.* [35] argue that residual networks implicitly represent an ensemble of shallow networks, thus allowing the architecture to learn a highly non-linear function and produce outstanding results.

Our model is based on the very deep ResNet-101 introduced by He *et al.* [6]. In contrast to the semantic segmentation model introduced by Chen *et al.* [15], we do not use parallel replications of ResNet-101 on different scales as it makes the network computationally expensive to train. Instead, we use several parallel convolutional layers of the same kernel size, but different dilation parameters. This way we transform the convolutional features into parallel segmentation maps of different receptive fields. Separate dilated convolutions allow us to enlarge the effective receptive field of the CNN. This context information improves the network’s understanding of text at different scales. Dilated convolutions do not increase the number of parameters, ensuring that the model remains easy to train. Lastly, the obtained parallel segmentation maps are fused together by element wise summation, providing the final segmentation map, which can then be used for word extraction.

3.4. Weighted Softmax Loss Function

A common loss function for training semantic segmentation networks is a pixelwise classification softmax loss. Such a function is appropriate for dense pixelwise labelling if there are many classes. For text localization, the pixelwise softmax loss tends to force the network to produce merged segmenations on the borders of words results such as the ones illustrated in Fig. 3.3. Post processing techniques are required to enhance the segmentation bounding boxes in order to use them for text recognition. In order to overcome this problem, a simple and efficient technique is introduced: instead of a binary text/non-text classification we define the notion of a border for each separate



Figure 3.2.: Word localization as semantic image segmentation. Individual words are trained to be split up using purely visual information. Left and middle-left: confidence levels as heatmaps for text and WordFence area. Mid-right: segmentation results. Far-right: word bounding box obtained from segmentation by connected component analysis.

word as a third class. The border acts as a penalization for training. The model is driven to surround each separate word with an artificial barrier, which greatly reduces the ease and computational cost of reading separate words. During inference, individual words are cleanly segmented from each other and can then be extracted using connected components analysis.

Since the number of text pixels in a text recognition dataset may not be balanced among labels and the vast majority of all pixels are simply background - networks tend to predict background everywhere. To solve this issue, we introduce a weighted normalization. The new loss function automatically penalizes predictions for pixels which form the majority of a given image and emphasizes pixels which are fewer in number. This makes the loss function well constrained for the task of text segmentation. Weight normalization is applied in two places: loss calculation and loss backpropagation. The normalization factors are calculated on the fly and are inversely proportional to the count of pixels of each class. The general algorithm of the weighted softmax loss function is shown in Algorithm 1.

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: $n_1, n_2, \dots, n_m \leftarrow CountsOfUniqueLabels(\mathbf{L})$ ▷ 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: $Backpropagate(loss, \frac{1}{n_1}, \frac{1}{n_2}, \dots, \frac{1}{n_m})$ ▷ loss backpropagation with normalization factors
-

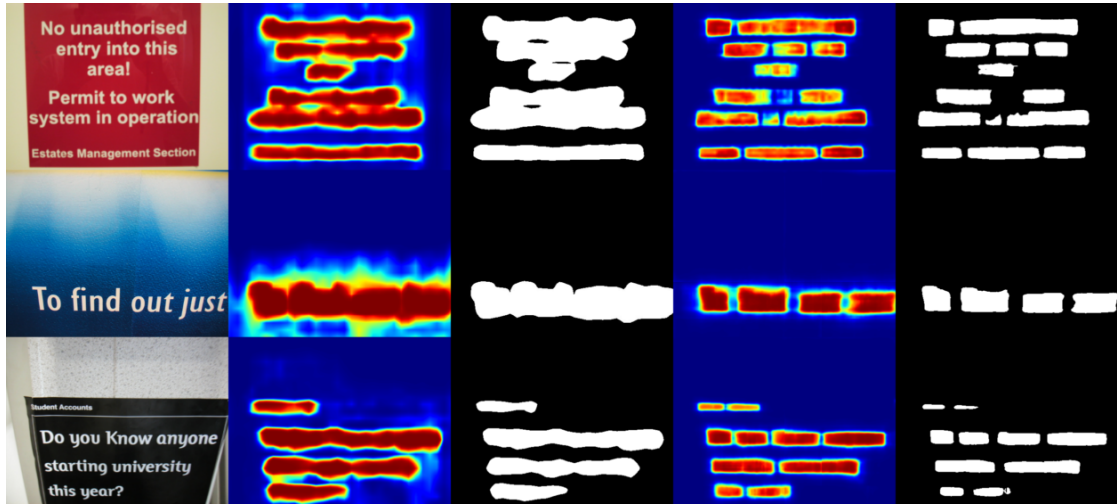


Figure 3.3.: Comparisons of segmentation with and without WordFence. The first column from the left shows the original images. The second and third show the text position belief map and the resulting segmentation, respectively. Last two columns show the belief map and the segmentation from our method. Localizing words without WordFence has a tendency of individual words bleeding over into each other, which causes difficulties to posterior word recognition.

4. Evaluation and Experiments

4.1. Datasets

Our model is evaluated on a number of different datasets (See Table 4.1). There is a large variety of quality in the datasets. Some are only used for training and others for evaluation. The datasets are summarized in Table 4.1.

The COCO-Text dataset tagged by Veit *et al.* [34] is based on the earlier MS-COCO dataset meant for object classification [16]. The full COCO-Text dataset consists of 20510 images with legible, machine labeled text in English. SynthText consists of 858750 natural images with synthetic text labels. The synthetic labels show a high level of sophistication with randomized locations of text, varying fonts, sizes and colors of text. ICDAR 2011 [27] and ICDAR 2013 [4] are common benchmark datasets from the International Conference of Document Analysis and Recognition. Street View Text dataset (SVT) [37] was harvested from Google Street View images, the images in this dataset exhibit high variability and low resolution making it a good metric for inspection, however the ground truth is often mislabeled resulting in low test accuracies. We train our model on MS-COCO, finetune on SynthText and then evaluate on ICDAR11, ICDAR13 and SVT for comparison with other state-of-the-art methods.

4.2. Word Detection Experiments

For the evaluation of our word detection results we use a PASCAL VOC style protocol where a proposal with intersection-over-union (IoU) ≥ 0.5 is considered a positive detection. PASCAL VOC is suitable for detecting individual words as it penalizes areas covering multiple words. The other common evaluation metric - DetEval [40] is better suited to textlines as it does not penalize merged words. PASCAL VOC evaluation protocol on the other hand is more suitable for individual word detection as it heavily

Label	Description	Mode	# Images
Synth	SynthText in the wild	Train	858750
COCO	COCO-Text	Train	20510
IC11	ICDAR 2011	Test	255
IC13	ICDAR 2013	Test	233
SVT	SVT	Test	250

Table 4.1.: Description of the text recognition datasets

Model	PASCAL VOC IoU = 0.5								
	ICDAR11			ICDAR13			SVT		
	P	R	F	P	R	F	P	R	F
Tian <i>et al.</i> [32]	0.89	0.79	0.84	0.93	0.83	0.88	-	-	-
Gupta <i>et al.</i> [5]	0.78	0.63	70.0	0.78	0.63	0.70	0.47	0.45	0.46
Jaderberg <i>et al.</i> [11]*	0.89	0.68	77.4	0.89	0.68	0.77	0.59	0.49	0.54
Gupta <i>et al.</i> [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

Table 4.2.: Comparisons with other methods of word detection. Precision, Recall and F-Score are reported. Recall maximization was necessary for obtaining good word detection results. Methods marked with * use a multistage false-positive filtering process to increase precision, the code was not published thus the results are not directly comparable with ours.

penalizes areas covering multiple words. The PASCAL VOC metric illustrates the real number of words that were detected in an image, and are immediately ready for the recognition stage.

Running the image inference at different scales produces different segmentation maps that need to be processed afterwards. When merging segmentations from different scales, the results will contain many duplicates and false positives, but recall will be high since true positives will likely have been found. We adopt a mechanism for merging segmentation maps of different scales before extracting the bounding boxes, while maintaining a high recall. We use a voting scheme to produce a final segmentation map. To do that we upscale all segmentation maps and find labels that correspond to maximal class probabilities in the segmentation maps. We extract the probability values for the found labels and sum them up on corresponding channels producing the map of summed maximum probabilities from different scales. The final segmentation is obtained by finding the labels with maximum probabilities on the combined map giving fewer false positives. Merge processing is fast and computationally cheap as it only needs to upsample images.

Table 4.2 shows the performance of our WDN model on the benchmark datasets. Although we did not obtain the highest precision, our model achieved a significant improvement in recall in comparison with previous state-of-the-art methods that do not incorporate filtering steps, generating a very small number of false positives (15 on average per image). On average we improved recall by 15% over the previous multi-scale detection method by Gupta *et al.* [5].

4.3. Timings

During test time the WDN model processes one image of four scales including word extraction in 2 seconds on a GPU (input size 513×513 px). The scales of images were chosen to be $\frac{1}{2}$, 1, $\frac{3}{2}$ and 2. So, from one average image at most 10 crops are generated

Label	Time	Model	Year	IC11	IC13
Jaderberg <i>et al.</i> [11]	7.00 (s)	Neumann <i>et al.</i> [20]	2013	0.45	-
Gupta <i>et al.</i> [5]	2.47 (s)	Jaderberg <i>et al.</i> [11]	2015	0.69	0.76
WDN	2.00 (s)	Gupta <i>et al.</i> [5]	2015	0.84	0.85
		WDN	2016	0.84	0.86

Table 4.3.: Comparison of word detection time (in seconds)

Table 4.4.: Evaluation of end-to-end word recognition on ICDAR 2011 and 2013 datasets. F-score is reported.

in order to run a multi-scale inference: one crop for resolutions $\frac{1}{2}$ and 1 and 4 crops for resolutions 1.5 and 2. In comparison with previous detection approaches (see Table 4.3) we gain a $0.25\times$ speed-up for the overall execution time for a single image as compared with Gupta *et al.* [5], which achieved the highest precision.

In contrast to works by Jaderberg *et al.* [11] and Gupta *et al.* [5] our approach generates an average of 15 detections per images eliminating the need of difficult and time consuming post-processing steps like random-forest classifier, CNN bounding box regression *etc.*

4.4. End-to-end Word Detection and Recognition

Using ideal, single-word proposals recognition accuracy can be as high as 98% [11]. In order to show the effectiveness and quality of proposals we integrate our model with a state-of-the-art recognition model by Shi *et al.* [28]. The recognition model consists of a CNN with an RNN component to recognize words of different length. Our word proposals are cropped out and evaluated with the recognition network.

We followed the evaluation protocol outlined by Wang *et al.* [36], where all word proposals that are three characters long or less or those that contain non-alphanumeric characters are ignored. An IoU overlap of 0.5 is required for a positive detection. The results for common recognition dataset are illustrated in Table 4.4. Our detection network achieves state-of-the-art recall rates - ensuring good candidate words. This combined with the recognition module obtains very accurate results for end-to-end word recognition. The network outperforms results by Jaderberg *et al.* [11] and is on par or better than Gupta *et al.* [5] while working in linear time.

5. Conclusion

5.1. Summary

In this paper we have presented a novel WordFence Detection Network. WDN relies on space between words to learn how to accurately split words using purely visual information, even for a wide variety of fonts, text sizes, scales, orientations and text languages. After segmenting an image proposal bounding boxes are extracted at multiple scales with very high detection recall. Lastly, end-to-end word recognition achieves state-of-the-art results with 84 % and 86 % F-Score on ICDAR11 and ICDAR13, respectively. We obtain such high end-to-end scores by leveraging the high quality proposals and high recall of the detection stage. Experimental results show that our approach achieves very competitive performance on ICDAR11 and ICDAR13 without utilizing any heuristics or knowledge based approaches.

The work done can be summarized into the following work steps:

- Analysis of available machine learning technologies and state-of-the-art
- Training network for text localization using semantic segmentation
- Adding WordFence areas to delineate individual words
- Adding pixelwise weighted softmax loss function as penalization for training
- Evaluation of the proposed solution

5.2. Dissemination

This project has been developed in partnership with the A*STAR Institute for Infocomm Research in Singapore. It has been submitted for review to the IEEE International Conference on Image Processing (ICIP) 2017 that will take place in Beijing, China in September 2017.

5.3. Problems Encountered

Some of the major problems that were encountered while developing the project include:

- Going from segmentations to bounding boxes
- Noisy detections and false positives

5.3.1. Segmentations to bounding boxes

We used connected components analysis [30] for transforming segmentations into bounding box ROIs. The problem with this approach is that it only yields rectangular results. This was not a huge problem because the ICDAR datasets generally only has horizontal text. It is foreseeable however that this limitation would not deal with text rotated at an angle successfully. Rotated text would require matching a best-fit rotated rectangle bounding box to the predicate connected region.

5.3.2. Noisy detections and false positives

It was difficult to filter out small false positive regions. This was an extremely challenging problem because of the delicate balance between precision and recall (see Table 4.2). Having a large number of proposals ensures that recall is high, because the true positives are more likely to be covered by at least one of the proposals. However, precision tends to be low when there are many false positives.

$$Precision = \frac{TruePositives}{(TruePositives + FalsePositives)} \quad (5.1)$$

$$Recall = \frac{TruePositives}{(TruePositives + FalseNegatives)} \quad (5.2)$$

Maintaining a high recall was central to our end-to-end word recognition framework. Knowing that we could still get a high F-Score in word recognition we decided to maintain high recall at the expense of precision in the detection stage.

5.4. Future Work

The WDN relies on purely visual information for delineating individual words. While, it is impressive that it is able to handle the large variety of possible fonts, character sizes and text kerning, it is not the same way that humans read. Humans, while reading text are able to utilize word semantics to understand where words begin, end or if they have spelling or other irregularities. In fact according to the aptly titled paper "Reading words with jumbled letters" by Rayner *et al.* [22], normal people can easily correct misspelled words while reading text, just based on approximate length and beginning and starting symbols. The combination of natural language processing and semantic understanding of words along with the visual information of the specific words and characters is needed for true understanding of text in natural scene images. Semantic models could be built into a text recognition network using RNNs or other memory mechanisms. Such a model could be trained to directly recognize text in an image, without intermediate computer vision steps and it would be a great step forward for artificial intelligence.

List of Acronyms

A*STAR	Singaporean Agency for Science, Technology and Research
CNN	Convolutional neural network
COCO Text	Text dataset extracted from MS-COCO
ConvNet	Convolutional network
FCN	Fully convolutional network
F-RCNN	Faster R-CNN
I ² R	A*STAR Institute for Infocomm Research
ICDAR	International Conference on Document Analysis and Recognition
ICIP	IEEE International Conference on Image Processing
IEEE	Institute of Electrical and Electronics Engineers
IoU	Intersection-over-union
MS-COCO	Microsoft Common Objects in Context
OCR	Optical character recognition
R-CNN	Regions with Convolutional Neural Network Features
ResNet	Residual network
RNN	Recurrent neural network
RoI	Region of interest
SOTA	State-of-the-art
SSD	Single Shot Decoder
SVT	Google Streetview Dataset
WDN	WordFence Detection Network
YOLO	You Only Look Once

Bibliography

- [1] J. C. Caicedo and S. Lazebnik. Active object localization with deep reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2488–2496, 2015.
- [2] B. Epshtein, E. Ofek, and Y. Wexler. Detecting text in natural scenes with stroke width transform. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2963–2970. IEEE, 2010.
- [3] S. Gidaris and N. Komodakis. Locnet: Improving localization accuracy for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 789–798, 2016.
- [4] L. Gomez and D. Karatzas. Multi-script text extraction from natural scenes. In *2013 12th International Conference on Document Analysis and Recognition*, pages 467–471. IEEE, 2013.
- [5] A. Gupta, A. Vedaldi, and A. Zisserman. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2315–2324, 2016.
- [6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- [7] T. He, W. Huang, Y. Qiao, and J. Yao. Accurate text localization in natural image with cascaded convolutional text network. *CoRR*, abs/1603.09423, 2016.
- [8] T. He, W. Huang, Y. Qiao, and J. Yao. Text-attentional convolutional neural network for scene text detection. *IEEE Transactions on Image Processing*, 25(6):2529–2541, 2016.
- [9] W. Huang, Y. Qiao, and X. Tang. Robust scene text detection with convolution neural network induced msr trees. In *European Conference on Computer Vision*, pages 497–511. Springer, 2014.
- [10] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. *CoRR*, abs/1406.2227, 2014.
- [11] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision*, 116(1):1–20, 2016.
- [12] M. Jaderberg, A. Vedaldi, and A. Zisserman. Deep features for text spotting. In *European conference on computer vision*, pages 512–528. Springer, 2014.
- [13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [14] Y. Li, K. He, J. Sun, et al. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems*, pages 379–387, 2016.
- [15] C. Liang-Chieh, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *International Conference on Learning Representations*, 2015.

- [16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.
- [17] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer, 2016.
- [18] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.
- [19] Y. Lu, T. Javidi, and S. Lazebnik. Adaptive object detection using adjacency and zoom prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2351–2359, 2016.
- [20] L. Neumann and J. Matas. Real-time scene text localization and recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3538–3545. IEEE, 2012.
- [21] E. Park and A. C. Berg. Learning to decompose for object detection and instance segmentation. *arXiv preprint arXiv:1511.06449*, 2015.
- [22] K. Rayner, S. J. White, R. L. Johnson, and S. P. Liversedge. Raeding wrods with jubmled lettres there is a cost. *Psychological science*, 17(3):192–193, 2006.
- [23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.
- [24] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.
- [25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.
- [26] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, abs/1312.6229, 2013.
- [27] A. Shahab, F. Shafait, and A. Dengel. Icdar 2011 robust reading competition challenge 2: Reading text in scene images. In *2011 international conference on document analysis and recognition*, pages 1491–1496. IEEE, 2011.
- [28] B. Shi, X. Bai, and C. Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.
- [29] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.
- [30] S. Suzuki et al. Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics, and Image Processing*, 30(1):32–46, 1985.
- [31] S. Tian, Y. Pan, C. Huang, S. Lu, K. Yu, and C. Lim Tan. Text flow: A unified text detection system in natural scene images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4651–4659, 2015.
- [32] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao. Detecting text in natural image with connectionist text proposal network. In *European Conference on Computer Vision*, pages 56–72. Springer, 2016.

- [33] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in neural information processing systems*, pages 1799–1807, 2014.
- [34] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. In *arXiv preprint arXiv:1601.07140*, 2016.
- [35] A. Veit, M. Wilber, and S. Belongie. Residual networks are exponential ensembles of relatively shallow networks. *arXiv preprint arXiv:1605.06431*, 2016.
- [36] K. Wang, B. Babenko, and S. Belongie. End-to-end scene text recognition. In *2011 International Conference on Computer Vision*, pages 1457–1464. IEEE, 2011.
- [37] K. Wang and S. Belongie. Word spotting in the wild. In *European Conference on Computer Vision*, pages 591–604. Springer, 2010.
- [38] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng. End-to-end text recognition with convolutional neural networks. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 3304–3308. IEEE, 2012.
- [39] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.
- [40] C. Wolf and J.-M. Jolion. Object count/area graphs for the evaluation of object detection and segmentation algorithms. *International Journal of Document Analysis and Recognition (IJDAR)*, 8(4):280–296, 2006.
- [41] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016.
- [42] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai. Multi-oriented text detection with fully convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4159–4167, 2016.
- [43] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1537, 2015.
- [44] Z. Zhong, L. Jin, S. Zhang, and Z. Feng. Deeptext: A unified framework for text proposal generation and text detection in natural images. *arXiv preprint arXiv:1605.07314*, 2016.
- [45] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision*, pages 391–405. Springer, 2014.

Appendices

A. Text Detection

This appendix provides selected sample detection results from the ICDAR11, ICDAR13 and SVT datasets. Precision, Recall and F-Score (overlapping IoU >0.5) as well as observations are reported for each image. The ground truth is highlighted in blue and the WDN bounding box is highlighted in green. Best viewed in color.

A.1. ICDAR 2011



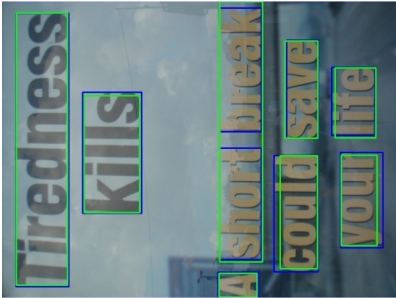

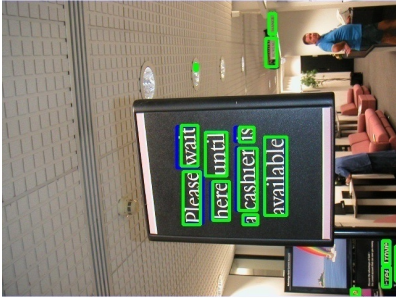

			
P=0.875, R=0.778, F=0.824 "short break" not split correctly	P=0.625, R=1.000, F=0.769 Some false positives	P=0.667, R=1.000, F=0.800	P=0.800, R=1.000, F=0.889

Table 1.: ICDAR11 Results Part 1

				P=1.000, R=1.000, F=1.000	P=1.000, R=0.333, F=0.500	P=0.250, R=1.000, F=0.400	P=0.778, R=0.636, F=0.700
	This is arguably one word	Pictorials detected as words	Words not split correctly				
				P=1.000, R=1.000, F=1.000	P=1.000, R=1.000, F=1.000	P=0.500, R=0.250, F=0.333	P=1.000, R=1.000, F=1.000
		Small text is harder to split					

Table 2.: ICDAR11 Results Part 2

A.2. ICDAR 2013

			
P=1.000, R=0.400, F=0.571	P=0.300, R=0.750, F=0.429	P=0.778, R=0.875, F=0.824	P=0.900, R=1.000, F=0.947
Challenging case - word is within another word	Too much word splitting, due to tricky font and spacing		

Table 3.: ICDAR13 Results Part 1

				P=1.000, R=0.889, F=0.941	P=0.833, R=0.714, F=0.769	P=1.000, R=1.000, F=1.000	P=1.000, R=0.933, F=0.966
				Some overlaps are <0.5 IoU	P=0.650, R=0.887, F=0.743	P=0.833, R=0.833, F=0.833	Training data did not have many numbers, network has trouble with them
P=0.000, R=0.000, F=0.000 All IoU overlaps are <0.5	P=0.800, R=1.000, F=0.889	P=0.650, R=0.887, F=0.743	P=0.833, R=0.833, F=0.833				

Table 4.: ICDAR13 Results Part 2

A.3. SVT





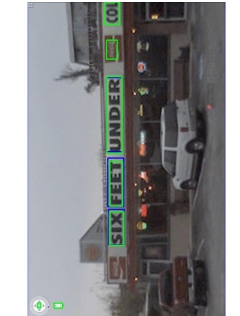


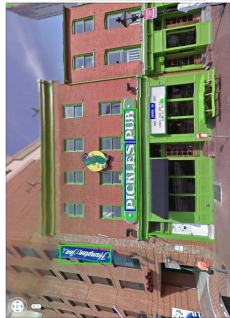

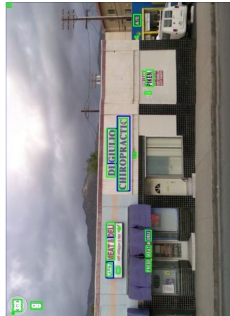
				
0.182, 0.500, 0.267	0.500, 1.000, 0.667	0.048, 0.250, 0.080	0.091, 0.500, 0.154	0.429, 1.000, 0.600
				
0.000, 0.000, 0.000	0.273, 0.500, 0.353	0.250, 0.250, 0.250	0.100, 1.000, 0.182	0.190, 0.800, 0.308

Table 5.: SVT Results: The SVT dataset ground truth is not well-labeled and the experimental results on this dataset do not necessarily correspond to a good localization network. However, a visual, qualitative inspection of the results shows that WDN detects and splits the majority of words present. It is also interesting to note that the Google Maps compass is often detected as a word - showing the same trend of detecting pictorials as in Table2