Total-Text: A Comprehensive Dataset for Scene Text Detection and Recognition (EXTENDED VERSION)

Chee Kheng Ch'ng Chee Seng Chan
Centre of Image & Signal Processing, Faculty of Computer Science & Info. Technology,
University of Malaya, Malaysia
chngcheekheng@siswa.um.edu.my, cs.chan@um.edu.my

***Abstract*—Text in curve orientation, despite being one of the common text orientations in real world environment, has close to zero existence in well received scene text datasets such as ICDAR'13 and MSRA-TD500. The main motivation of Total-Text is to fill this gap and facilitate a new research direction for the scene text community. On top of conventional horizontal and multi-oriented text, it features curved-oriented text. Total-Text is highly diversified in orientations, more than half of its images have a combination of more than two orientations. Recently, a new breed of solutions that casted text detection as a segmentation problem has demonstrated their effectiveness against multi-oriented text. In order to evaluate its robustness against curved text, we fine-tuned DeconvNet and benchmark it on Total-Text. Total-Text with its annotation is available at https://github.com/cs-chan/Total-Text-Dataset.**

***Keywords*-Scene text dataset; Curve-oriented text; Segmentation-based text detection**

I. Introduction

Scene text detection is one of the active computer vision topics due to the growing demands of applications such as multimedia retrieval, industrial automation, assisting device for vision-impaired people, etc. Given a natural scene image, the goal of text detection is to determine the existence of text, and return the location if it is present.

Well known public datasets such as ICDAR'03, '11, '13 [1] (term as ICDARs from here onwards), and MSRA-TD500 [2] have played a significance role in initiating the momentum of scene text related research. One similarity in all the images of ICDARs is that all the texts are in horizontal orientation [12]. Such observation has inspired researchers to incorporate horizontal assumption [3]–[7] in solving the scene text detection problem. In 2012, Yao *et al*. [2] introduced a new scene text dataset, namely MSRA-TD500, that challenged the community with texts arranged in multiple orientations. The popularity of it in turn defined the convention of 'multi-oriented' texts. However, a closer look into the MSRA-TD500 dataset revealed that most, if not all the texts are still arranged in a straight line manner as to ICDARs (more details in Section III). Curved-oriented texts(term as curved text from here onwards), despite its commonness, are missing from the context of study. To the best of our knowledge, CUTE80 [8] is the only available scene text dataset to-date with curved text. However, its scale is too small with only 80 images and it has very minimal scene diversity.

Figure 1: Annotation details of Total-Text, including transcription, polygon-shaped and rectangular bounding box vertices, orientations, *care* and *do not care* regions, and binary mask.

Without the motivation of a proper dataset, effort in solving the curved text detection problem is rarely seen. This phenomenon brings us to our primary contribution of this paper: Total-Text, a scene text dataset collected with curved text in mind, filling the gap in scene text datasets in terms of text orientations. It has 1,555 scene images, 9,330 annotated words with 3 different text orientations including horizontal, multi-oriented, and curved text.

Orientation assumption is commonly seen in text detection algorithms. We believe that the heuristic design to cater different types of text orientations hold back the generalization of text detecting system against texts in the real world with unconstrained orientations. Recent works [9]–[11] have started to cast text detection as a semantic segmentation problem, and achieved state-of-the-art results in ICDAR'11, '13 and MSRA-TD500 datasets. They have reported successful detection of curved text as well. He et

Figure 2: Curved text is commonly seen in real world scenery.

Figure 3: 1st row: Examples from ICDAR 2013, ICDAR2015 and MSRA-TD500; 2nd row: Slightly curved to extremely curved text examples from the Total-Text.

(a) Yin et. al. [22] (red bounding box) and Huang *et al*. [6] (blue bounding box)

(b) Shi *et al*. [7]

Figure 4: These show that the current state-of-the-art solutions could not detect curved text effectively.

al.[3] system in particular has no orientation assumption and hueristic grouping mechanism. This bring us to the secondary contribution of this paper, we looked into this new solution and revealed how it handle multiple oriented text in natural scene.

II. Related Works

This section will discuss closely related works, specifically scene text datasets and text detection system. For completeness, readers are recommended to read [12].

A. Scene Text Datasets

ICDARs [1] has three variants. ICDAR'03 started out with 509 camera taken scene text images. All the scene texts in the dataset appear in horizontal orientation. In ICDAR'11, the total number of images were reduced to 484 to eliminate duplication in the previous version. ICDAR'13 further trimmed down the 2011 version to 462 images in total. Improvement was done to increase its text categories and tasks. In ICDAR'13, there are 462 images of horizontal English texts. Recently, ICDAR launched a new challenge [13] named as the 'Incidental Scene Text' (also known as the ICDAR'15), which is based on 1670 images captured with wearable devices. It is more challenging than previous datasets as it has included text with arbitrary orientation and most of them are out of focus.

MSRA-TD500 [2] was introduced in 2012 to address the lack of arbitrary orientated text in scene text datasets. It has 300 training and 200 testing images; annotated with minimum area rectangle.

COCO-text [14] was released in the early 2016, and is the largest scene text dataset to-date with 63,686 images and 173,589 labeled text regions. This large scale dataset contains all variety of text orientations: horizontal, arbitrary and curved. However, it used the axis oriented rectangle as groundtruth, which seems to be applicable only to horizontal and vertical texts.

CUTE80 [8] is the only curved text dataset available in public to the best of our knowledge. It has only 80 images and limited sceneries.

B. Scene Text Detection:

Scene text detection has seen significant progress after the seminal work by Epshtein *et al*. [15] and Neumann and Matas [16]. In the former, Stroke Width Transform (SWT) was proposed to detect text. This method considered similar stroke widths to group text components and studied the component properties to classify them. In the latter, Maximally Stable Extremal Regions (MSER) was exploited to extract text components. They used geometrical properties of the components and a classifier to detect text. Both represent character better than all other feature extractors like color, edge, texture and etc. Upon picking up potential character candidates, these connected components based algorithms typically go through text line generation, candidates filtering and segmentation as pointed out by this survey [12].

As to many other computer vision tasks, the incorporation of Convolutional Neural Network (CNN) in localizing text is a very active research at the moment. Huang *et al*. [6] trained a character classifier to examine components generated by MSER, with the objective of improving the robustness of feature extraction process. Alongside this work, [17], [18] also trained a CNN to classify text components from non-text. This line of work demonstrated the high discriminative power of CNN as a feature extractor. However, interestingly, Zhang *et al*. [9] argued that leveraging on CNN as a character detector has restricted the CNN's potential due to the local nature of characters. Zhang *et al*. trained two Fully Convolutional Networks (FCN) [19]: 1) A Text-Block FCN that considers both local and global contextual info at the same time to identify text regions in an image, 2) Character-Centroid FCN to eliminate false text line candidates. However, text line generation, which plays a key role in grouping characters into a word, did not receive much benefit from the robust CNN. While most of the algorithms [9], [18] handcrafted the text line generation process, He *et al*. [10] trained a FCN to infer text line candidates. By cascading a text region and a text line using supervised FCN, Cascaded Convolution Text Network (CCTN) achieved generalization in terms of text orientations, and is one of the best performing system in both horizontal and abritrary oriented scene text datasets: ICDAR 2013 and MSRA-TD500.

III. Total-Text Dataset

This section will discuss a) the motivation of collecting Total-Text; b) observation made on horizontal, multi-oriented, and curved text; c) orientation assumption aspect in the current state-of-the-art algorithms, and d) different aspects and statistics of Total-Text.

A. Dataset Attributes

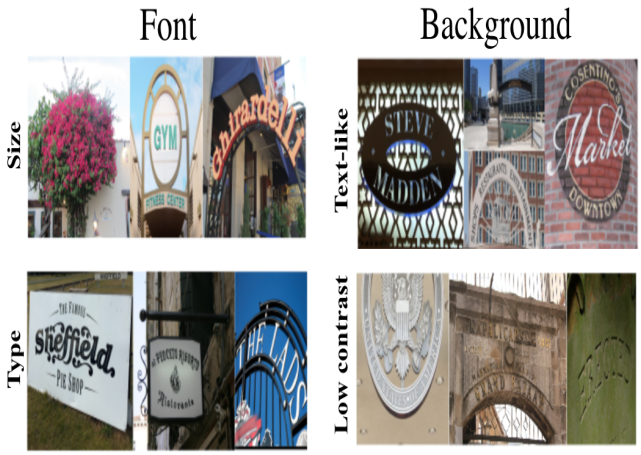
Curved text is an overlooked problem. The effort of collecting this dataset is motivated by the missing of curved text in existing scene text datasets. Curved text can be easily found in real life scenes such as: business logos, signs, entrances etc as depicted in Fig. 7d, surprisingly such data has close to zero existence in the current datasets [1], [2], [13]. The most popular scene text dataset over the decade, ICDARs have only horizontal text [12]. Consequently, vast majority of algorithms assume text linearity to tackle the problem effectively. As a result of overwhelming attention, performances of text detections in ICDARs are saturated at quite a high point (0.9 in terms of f-score). Meanwhile, multi-oriented text also received a certain amount of attention from this community. MSRA-TD500 is a well known dataset that introduced this challenge to the field. Algorithms like [9], [20] were designed to cater multi-oriented text. To the best of our knowledge, scene text detection algorithms designed for curved orientation [8] in consideration is relatively unpopular. We believe that the lack of such dataset is the obvious reason why the community has overlooked it. Hence, we propose Total-Text with 4,265 curved text out of 9,330 total text instances, hoping to spur an interest in the community to address curved text.

Curved text observation. Geometrically speaking, a straight line has no angle variation along the line, and thus can be described as a linear function, $y = mx + c$. *A curved line is not a straight line*. It is free of angle variation restriction throughout the line. Shifting to the scene text perspective, we observed that horizontal oriented text or word is a series of characters that can be connected by a straight line; their bottom alignment in particular for most cases. At the same time, multi-oriented text, in scene text convention, can also be connected by a straight line, given an offset with respect to a horizontal line. Meanwhile, characters a in curved word will not have unified angle offset, in which deemed to fit a polynomial line in text level (refer to Fig. 3 for image examples). In our dataset collection, we found out that curved text in natural images could vary from slightly curved to extremely curved. Also, it is not surprising to find that most of them are in the shape of a symmetric arc due to the symmetrical preferences in human vision [21].

Orientation assumption. We observed that orientation assumption is a must in a lot of algorithms [3]–[6], [9], [20]. We took a closer look into the orientation assumption aspect of existing text detection algorithms and see how it fits into the observation we have made on the curved text. We mainly focused on systems in which the authors claimed to have multi-oriented text detection capability and reported their results on MSRA-TD500. Zhang *et al*. [9] first used the FCN to create a saliency map and generate text blocks. Consequently, the system draw a straight line from the middle point of the generated text blocks, aiming to hit as many character components as possible; the straight line with the angle offset that hit the most text blocks will be considered as text line for the subsequent step. We believe that such mechanism would not work in our dataset, as a straight line would miss the polynomial nature of curved

Figure 5: Comparison between conventional rectangular bounding box (red colour) and the proposed polygon-shaped bounding region (green colour) in Total-Text. Polygon-shaped appeared to be the better candidate for groundtruth.

(a) Various text orientations (from left to right).
Top (One orientation): HC; VC; Cir and W.
Middle (Two orientations): Cir+H; MO+HC; W+H.
Bottom (Three orientations): H+MO+VC; H+MO+HC; H+MO+Cir



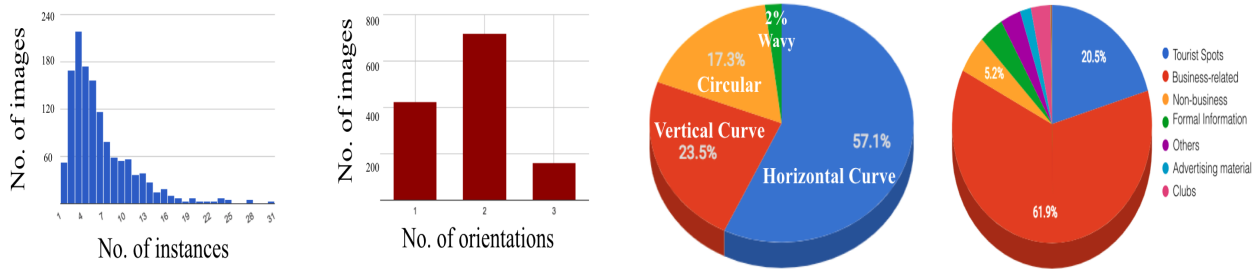
(b) Various text fonts and image backgrounds

Figure 6: Total-Text dataset is challenging due to its highly diversified orientation compositions and scenery.
Legends: H=horizontal, MO=multi-oriented, HC=horizontal curve, VC=vertical curve, Cir=circular and W=Wavy.

text. [20] focused on the text candidate construction part to detect multi-oriented text. Their algorithm will first clusters character pairs with consistent orientation or perspective view into the same group. As we can see in Fig. 3 (second row, second and third image specifically), characters in a single curved word could have multiple variations in terms of orientation. In fact, both of these algorithms, along with [7], have reported their failure on the same curved text images in MSRA-TD500 as illustrated in Fig. 4b. It is worth to note that MSRA-TD500 has only 2 curved text instances in the entire dataset. Last but not least, we ran [22] and [6] on several images of Total-Text, results can be seen in Fig. 4.

Focused scene text as a start. Two of the latest scene text datasets, COCO-text and ICDAR 2015 emerged to challenge current algorithms with incidental images. For example, scene images in the ICDAR 2015 [13] were captured without prior effort in positioning the text in it. Although it was not mentioned explicitly, one can deduce the emergence of these datasets are possibly due to: i) Performances of various algorithms on previous ICDARs dataset have saturated at a rather high point, hence a new dataset with higher level of complexity is deem required, ii) Well focused scene text are not likely to be captured by devices in real world scenarios. While the work done in curved text detection is considerably rare, we believe that it is at its infant stage. Inspired by the improvement in scene text detection and recognition brought by focused scene text datasets, notably ICDARs, and MSRA-TD500, we believe that focused scene text instead of incidental scene text is more appropriate to kick start related research work.

Tighter groundtruth is better. ICDAR 2015 employed quadrilaterals in its annotation to cater perspective distorted text [13]. However, COCO-text used rectangular bounding boxes [14] like ICDAR 2013, which we think is a poor choice considering the text orientation variations in it. Fig. 15 illustrates the downside of such bounding box annotation. Text regions cover much of the background which is not an ideal groundtruth for both evaluation and training. In Total-Text, we annotated the text region with polygon shapes that fits tightly, and the groundtruth is provided in polygon



(a) Text instances per image (b) Text orientations per image (c) Curve variations (d) Occurrence of curved text

Figure 7: Statistics of Total-Text dataset

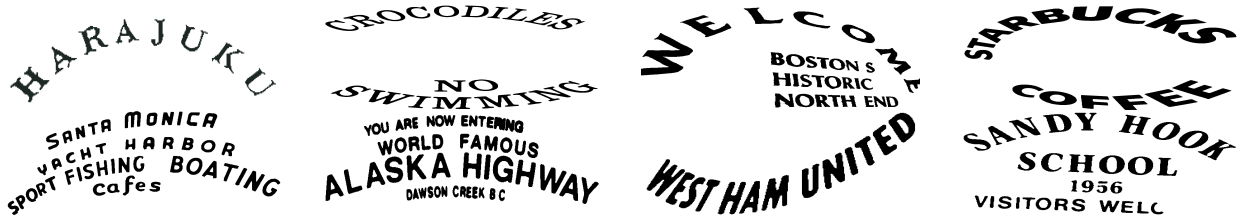


Figure 8: Examples of pixel-level annotation (cropped) in Total-Text.

vertices format.

Evaluation Protocol. Like ICDARs datasets [12], Total-Text uses DetEval [23]. We did a modication to the minimum intersection area calculation stage to handle our polygon-shaped groundtruth. The evaluation protocol will be made available as well.

Annotation Details. Groundtruth in the Total-Text is annotated in word level granularity. Adopted from the COCO-text, word level texts are *uninterrupted sequence of characters separated by a space*. As mentioned, Total-Text uses polygon shapes to bind groundtruth words tightly. Apart from that, we also included rectangular bounding box annotation considering most of the current algorithms generate rectangule bounding box outputs. However, it is not an accurate representation as a big chunk of background area is included due to the nature of curved text. Therefore, we do not encourage the usage of rectangular bounding box in our dataset. Total-Text considers only English characters in natural images; other languages, digital watermarks and unreadable texts are labelled as *do not care* in the groundtruth. *Do not care* area picks up by algorithms should be filtered out before evaluating its performance. Groundtruth for word recognition is also provided along with its spatial coordinates. In addition, orientation of every instances were annotated for modularity convenience. For example, if one prefer to evaluate curved text detection ability only, one could leverage this annotation to filter out intances with other orientations. Last but not least, Total-Text also comes with binary mask groundtruth to cater the recent requirements [9]–[11]. Fig. 1 illustrates all the aforementioned annotation details apart from the pixel-level annotation, which is illustrated in Fig. 8. Considering the scale of this dataset is manageable, authors of this paper annotated the entire dataset manually and cross checked with another 3 laboratory members.

B. Dataset Statistics

This subsection will discuss the statistics of Total-Text. All of the comparisons are made against ICDAR 2013 and MSRA-TD500, as they are the most common benchmark for horizontal and multi-oriented focused scene text respectively. Total-Text is split into two groups, training and testing set with 1255 and 300 images, respectively.

Strength in numbers. Fig. 7 shows a series of statistics information of the Total-Text. It has a total of 9330 annotated texts, 6 instances per image in average. More than half of the images in Total-Text have 2 different orientations and above, yielding 1.8 orientations per image on average. Both numbers ranked first against its competitors [12], showing the complexity of Total-Text. Apart from these solid numbers, the dataset was also collected with quality in mind, including scene complexity such as text-like and low contrast background, different font types and sizes, etc, image examples in Fig. 6b.

Orientation diversity. Approximate by half of the text instances are curved, and the other half is split almost equally between horizontal and multi-oriented. Curve text has its own variation too. Based on our observation, we classified them as horizontal curved, vertical curved, circular, and wavy (refer to 6a for image example). Their composition in the dataset can be seen in 7c. Although all the images were collected with curved text in mind, other orientations still occupy half of the total instances. A closer look into the

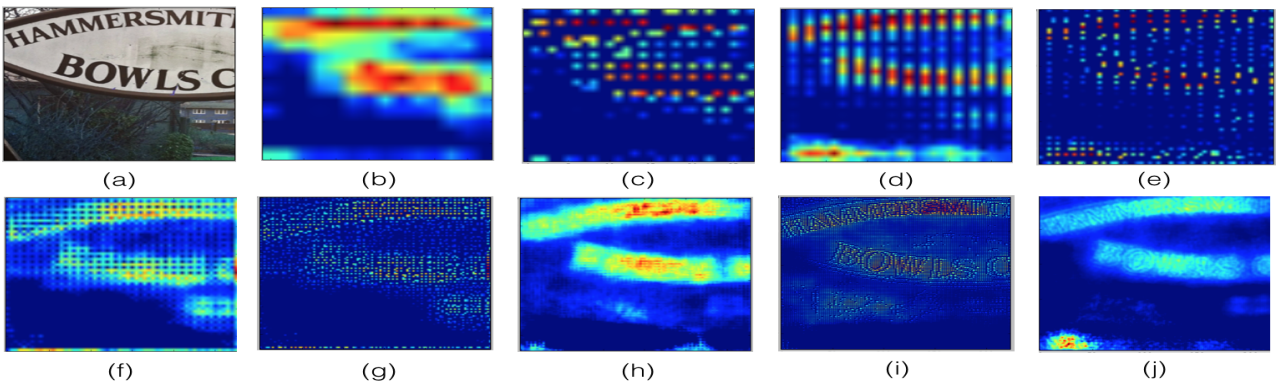


Figure 9: Visualization of the activations in deconvolution network. The activation maps from top left to bottom right correspond to the output maps from lower to higher layers in the deconvolution network. We select the most representative activation in each layer for effective visualization. (a) Input image; (b) the last 14×14 deconvolutional layer; (c) the 28×28 unpooling layer; (d) the last 28×28 deconvolutional layer; (e) the 56×56 unpooling layer; (f) the last 56×56 deconvolutional layer; (g) the 112×112 unpooling layer; (h) the last 112×112 deconvolutional layer; (i) the 224×224 unpooling layer and (j) the last 224×224 deconvolutional layer.

dataset shows that curved text usually appears with either horizontal or multi-oriented texts. The mixture of orientations in an image, challenges text detection algorithms to achieve robustness and generalization in terms of text orientations.

Scene diversity. In comparison to CUTE80 (the only publicly available curved text dataset), which majority of the images are football jerseys, Total-Text is much more diversified. Fig. 7d shows where curved text usually appears. Business related places like restaurant (*i.e.*, Nandos, Starbucks), company branding logos, and merchant stores take up of 61.2% of the curved text instances. Tourist spots such as park (*i.e.*, Beverly Hills in America), museums and landmarks (*i.e.*, Harajuku in Japan) occupy 21.1%. Fig. 2 illustrates these examples.

IV. Semantic Segmentation for Text Detection

Inspired by the success of FCN in the semantic segmentation problem, [9]–[11] casted text detection as a segmentation problem, and achieved state-of-the-art results. While most of the conventional algorithms failed in detecting curved text, their algorithms have shown successful results in limited number of examples due to the lack of available benchmark. The fact that [10] achieved good results without any heuristic grouping rules where most of the other algorithms need, intrigued us to look into this new breed of solution. We fine-tuned DeconvNet [24] and evaluated it on Total-Text, following section will discuss our findings.

A. DeconvNet

We select DeconvNet [24] as our investigation tool due to two reasons: 1) it achieved state-of-the-art results in semantic segmentation on Pascal VOC dataset and 2) Multiple deconvolutional layers in the DeconvNet allow us to observe the deviation finely. The scope of this paper is not proposing a new solution to solve the curved text problem, hence we merely convert and fine-tune the network to localize texts. For complete understanding, readers are encouraged to read [24].

Conversion. The last convolution layer of the original DeconvNet has 21 layers for 20 classes in the PASCAL VOC benchmark [25] and one background class. In this paper, we reduced it to two layers, representing text and non-text. Then, we fine-tuned the pre-trained model provided by Noh *et al*. [24] with one step training process instead of two as discussed in the original paper. Apart from these and the training data, all other training implementations were consistent with the original paper.

Training Data. Considering the depth of DeconvNet (*i.e.*, 29 convolutional layers and 252M parameters), we pre-trained it using the largest scene text dataset, COCO-text [14]. Images in the COCO-text were categorized into legible and illegible text, where we trained our network only on the legible text as it closely resemble our dataset. Similar to [9], [10], we first generated the binary mask with 1 indicating text region and 0 for background. Approximately 15k of training data were cropped into 256x256 patches to cater the receptive field of the DeconvNet. Patches with less than 10% text regions were eliminated to prevent overwhelming amount of non-text data. Roughly 200k and 80k patches of training and validation data were generated, respectively. We augmented the data in parallel to the training with horizontal flipping and random cropping (into 224x224).

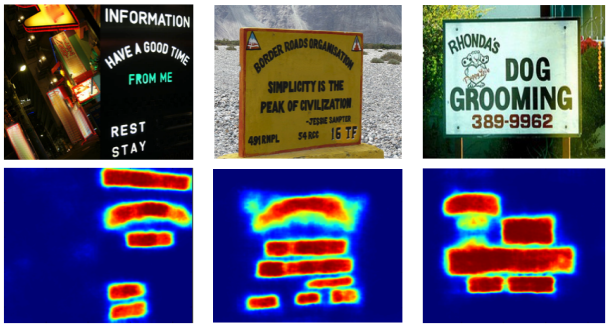


Figure 10: Successful examples of DeconvNet.

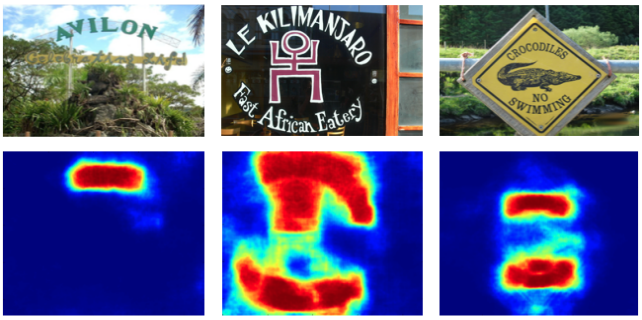


Figure 11: Failure examples of DeconvNet.

B. Experiments

Inference. The inference process was kept to be as simple as possible. We resized input images to 224×224, then forward propagated them through the DeconvNet. To generate final detection result, the saliency map was binarized using a threshold of 0.5, followed by connected component analysis to group 1s (text) pixels and bound them tightly with polygons.

Results. The outcomes were evaluated using our evaluation protocol and listed in Table I. As we went through each of the output saliency maps, we found two consistent roots that cause such unsatisfactory results: 1) The network is not robust enough for challenging backgrounds such as texts attached on repeated patterns such as bricks, gate, wall, etc.; 2) Multiple word candidates were grouped as one. Fig. 11 illustrates some failure examples. We suspect the robustness of the network was affected by its training data. Such loosely bounded training data with background regions labelled as 'text' could have impacted the training process to a certain extend. Meanwhile, producing word line level output is commonly seen in text detection algorithms, we lack of a segmentation process to separate them into words level.

Deeper look into the network. As mentioned before, our primary intention were to investigate the performance of DeconvNet on text with all sorts of orientations. With no orientation assumption or any heuristic grouping mechanism in the design, we managed to find candidates across texts

Table I: Evaluation of DeconvNet on Total-Text.

Dataset	Recall	Precision	F-score
Total-Text	0.33	0.40	0.36

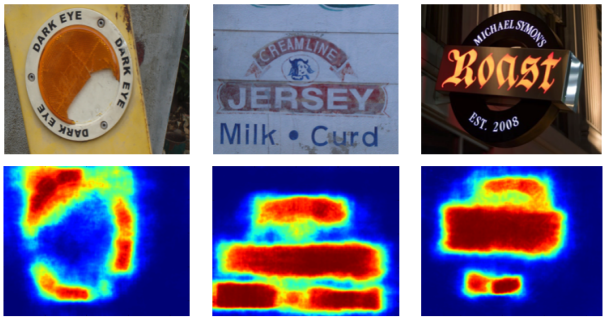


Figure 12: Examples of DeconvNet with lower confidence at both end of the curved text.

with all orientations as illustrated in Fig. 10. We were curious on how and what exactly happened across the deconvolution network. So, we cropped a specific patch of an original image that consists of curved text, forward propagated through the network, and observed the feature maps in several layers of the deconvolution network. As we can see in Fig. 9, at the lower layers, we can notice which part of the feature map is highly activated. As the layers proceed, finer details emerged, enriching the region of interest to an extend that we can recognized the characters in it.

Spatial resolution of feature maps is crucial. Text detection systems like [9], [10] adopted FCN and skip connections in their Convolutional Network. Such design element perserves spatial resolution of feature maps, and in turn provides better contextual information for their pixel-wise prediction task. Similarly, DeconvNet uses a combination of both unpooling layers and learn-able upsampling convolution filters to infer bigger feature maps layer after layer. As we can see in Fig. 10, such saliency map is high in resolution, depicts the actual shape or orientation of the detected text region. Minimal post-processing steps are required to retrieve text candidates from it.

Text line supervision is an interesting step forward. Fig. 12 illustrates several examples where the network is not confident about the shape of the curved text regions. We believe that it could be improved with text line supervision leveraged in [10]. This can be noticed in [10], where the work showed their results without the FTN, its performance droped from 0.84 to 0.5 in terms of F-score.

V. Conclusion

This paper introduces a comprehensive scene text dataset, Total-Text, featuring the missing element in current scene

text datasets - curved text. We believe that curved text should be included as part of the 'multi-oriented' text detection problem. While it is under research at the moment, we hope the availability of Total-Text could change the scene. We fine-tuned and analyzed how DeconvNet responds to curved text. Spatial resolution of feature maps and contextual information appeared to be crucial in segmentation based methods. Such methods are capable of predicting text regions in all sorts of orientations without hard-coded rules. Inspired by this observation, we plan to explore this area further with the aim of designing a scene text detect that is effective against multi-oriented text.

ACKNOWLEDGMENT

This work is partly supported by Postgraduate Research Grant (PPP) - PG350-2016A, from University of Malaya. The Titan-X GPU used by this research was donated by NVIDIA Corporation. We would also like to express our gratitude towards Jia Huei Tan, Yang Loong Chang and Yuen Peng Loh for Total-Text image collection and annotation.

REFERENCES

[1] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, and L. P. de las Heras, "Icdar 2013 robust reading competition," in *ICDAR*, 2013.

[2] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *CVPR*, 2012.

[3] Z. Zhang, W. Shen, C. Yao, and X. Bai, "Symmetry-based text line detection in natural scenes," in *CVPR*, 2015.

[4] W. Huang, Z. Lin, J. Yang, and J. Wang, "Text localization in natural images using stroke feature transform and text covariance descriptors," in *ICCV*, 2013.

[5] L. Neumann and J. Matas, "Scene text localization and recognition with oriented stroke detection," in *ICCV*, 2013.

[6] W. Huang, Y. Qiao, and X. Tang, "Robust scene text detection with convolution neural network induced mser trees," in *ECCV*, 2014.

[7] B. Shi, X. Bai, and S. Belongie, "Detecting oriented text in natural images by linking segments," in *CVPR*, 2017.

[8] A. Risnumawan, P. Shivakumara, C. S. Chan, and C. L. Tan, "A robust arbitrary text detection system for natural scene images," *Expert Systems with Applications*, vol. 41, no. 18, pp. 8027–8048, 2014.

[9] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai, "Multi-oriented text detection with fully convolutional networks," in *CVPR*, 2016.

[10] T. He, W. Huang, Y. Qiao, and J. Yao, "Accurate text localization in natural image with cascaded convolutional text network," *arXiv preprint arXiv:1603.09423*, 2016.

[11] C. Yao, X. Bai, N. Sang, X. Zhou, S. Zhou, and Z. Cao, "Scene text detection via holistic, multi-channel prediction," *arXiv preprint arXiv:1606.09002*, 2016.

[12] Q. Ye and D. Doermann, "Text detection and recognition in images and video : a survey," *T-PAMI*, vol. 37, no. 7, pp. 1–20, 2014.

[13] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu *et al.*, "Icdar 2015 competition on robust reading," in *ICDAR*, 2015.

[14] V. Andreas, M. Tomas, N. Lukas, M. Jiri, and B. Serge, "Coco-text: Dataset and benchmark for text detection and recognition in natural images," *arXiv preprint arXiv:1601.07140*, 2016.

[15] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *CVPR*, 2010.

[16] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761–767, 2004.

[17] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng, "End-to-end text recognition with convolutional neural networks," in *ICPR*, 2012.

[18] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep features for text spotting," in *ECCV*, 2014.

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.

[20] X.-C. Yin, W.-Y. Pei, J. Zhang, and H.-W. Hao, "Multi-orientation scene text detection with adaptive clustering," *T-PAMI*, vol. 37, no. 9, pp. 1930–1937, 2015.

[21] R. B. Adams, *The Science of Social Vision: The Science of Social Vision*. Oxford University Press, 2011, vol. 7.

[22] X.-C. Yin, X. Yin, K. Huang, and H.-W. Hao, "Robust text detection in natural scene images," *T-PAMI*, vol. 36, no. 5, pp. 970–983, 2014.

[23] C. Wolf and J.-M. Jolion, "Object count/area graphs for the evaluation of object detection and segmentation algorithms," *IJDAR*, vol. 8, no. 4, pp. 280–296, 2006.

[24] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *ICCV*, 2015.

[25] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

VI. Appendix

Figure 13 illustrates Total-Text dataset has very challenging attributes of real world scenery. For example, perspective distortion (Fig. 13a); variation in font types (Fig. 13b); variation in font sizes (Fig. 13c); background with text-like characteristics such as bricks, trees etc. (Fig. 13d); uneven lighting (Fig. 13e) and low contrast between text and background (Fig. 13f).

A. Text Orientations

Figure 14 shows examples with multiple text orientations. From unified orientation in Figure 14a to two orientations in Figure 14b-14c, to images with all sort of orientations in Figure 14d.

B. Different Groundtruth Evaluations

Table II and Figure 15 show the comparison between two different groundtruth in terms of *Precison* and *Recall*. Note that, Green is the detected text using the DeconvNet algorithm [24], Red is the groundtruth generated using the

(a) Perspective distorted examples.

(b) Different font type examples.

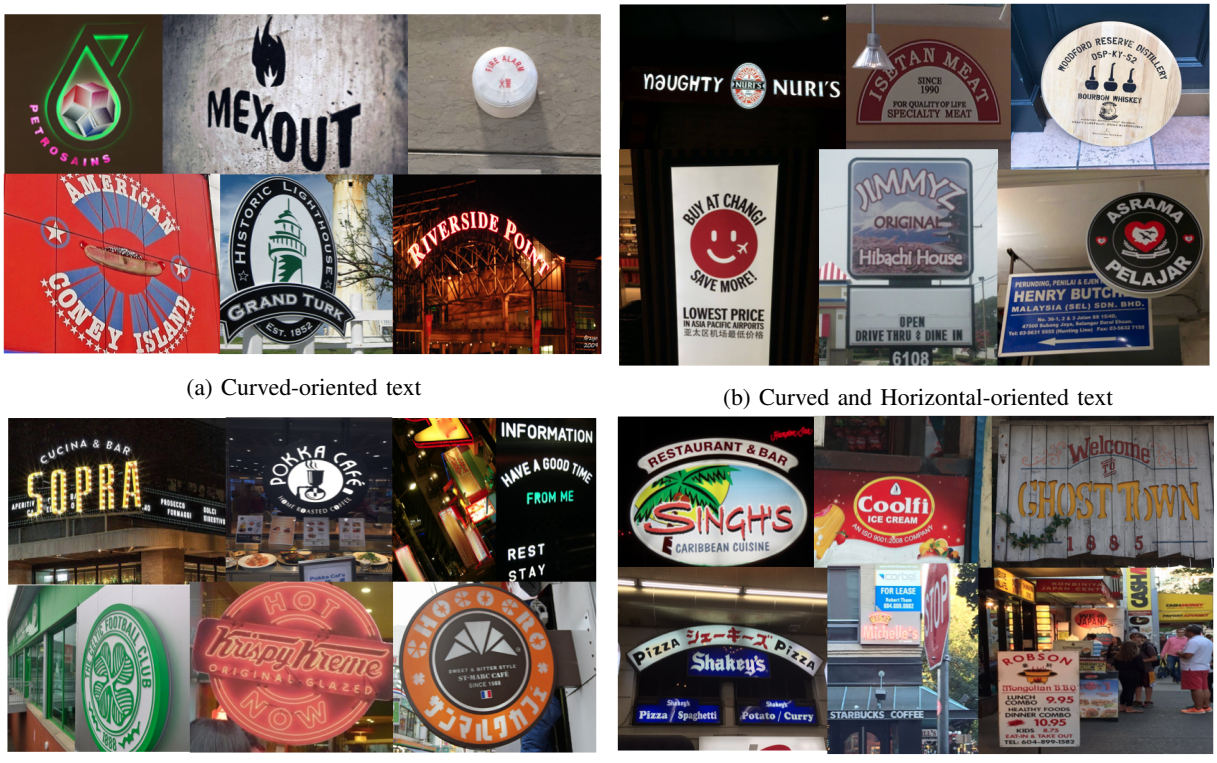
(c) Different font size examples.

(d) Complex background examples.

(e) Uneven lighting examples.

(f) Low contrast examples.

Figure 13: Challenging examples in the Total-Text dataset



(a) Curved-oriented text

(b) Curved and Horizontal-oriented text

(c) Curved and Multi-oriented text

(d) Curved and Horizontal and Multi-oriented text

Figure 14: Different text orientations in the Total-Text dataset

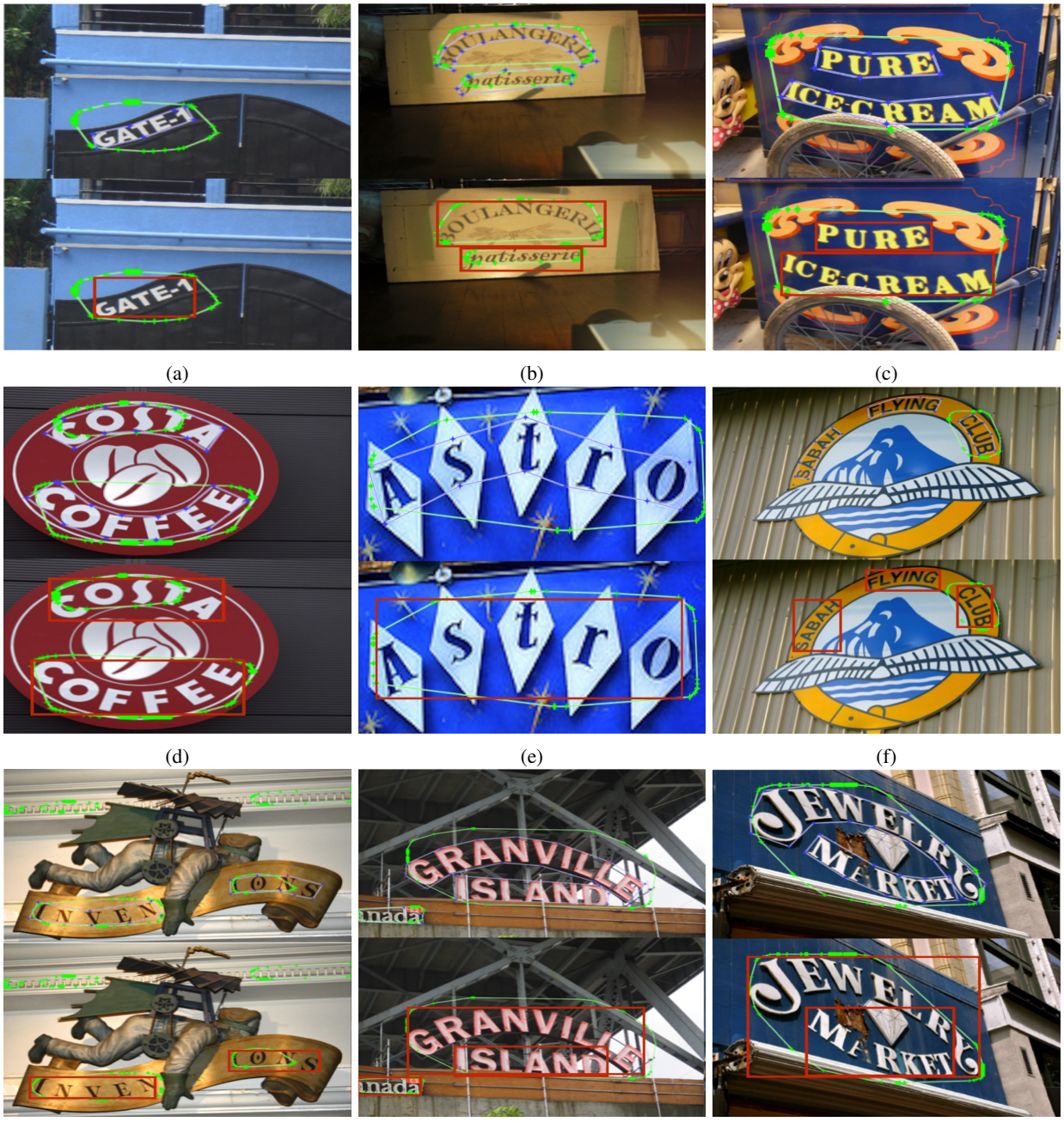
Figure	Matched Ground Truth	Polygon-shaped		Rectangle-shaped	
		Precision	Recall	Precision	Recall
15a	'GATE-1'	0.27	1	0.69	1
15b	'BOULANGERLE'	0.4	0.95	0.97	0.79
	'patisserie'	0.93	0.74	0.96	0.56
15c	'PURE'	0.13	1	0.16	1
	'ICE-CREAM'	0.23	0.99	0.44	0.98
15d	'COSTA'	0.86	0.64	0.95	0.43
	'COFFEE'	0.5	0.98	0.91	0.91
15e	'Astro'	0.48	0.98	0.9	0.93
15f	'CLUB'	0.43	1	0.73	0.94
15g	'INVEN'	0.65	0.98	0.89	0.91
	'IONS'	0.88	0.57	0.94	0.45
15h	'GRANVILLE'	0.37	0.98	0.86	0.93
	'ISLAND'	0.19	0.99	0.25	0.98
	'anada'	0.86	0.91	0.99	0.84
15i	'JEWELRY'	0.39	0.99	0.95	0.73
	'MARKET'	0.18	1	0.48	0.945

Table II: Evaluation results with different groundtruth format. Our proposed polygon-shaped groundtruth, provided alongside Total-Text bounds text regions tightly and hence provide a more accurate evaluation result.

conventional rectangular box and Blue is the groundtruth generated using our proposed polygon-shape. Figure 15a-15c show examples of the detected text (in green region) has higher precision score if we choose to employ the conventional rectangle-shaped groundtruth (red in color). Figure 15d-15i illustrate several examples of the detected text (in green region) have lower recall and precision because it misses a large intersection area with the groundtruth regions.

C. Groundtruth Examples

Figure 16 and 17 depict the annotation details of Total-Text dataset. Every images were annotated into four attributes: 1) spatial locations, 2) transcript, 3) orientation of each text instances, and 4) binary mask with annotated region as 1(white), background as 0(black).



(a) (b) (c)

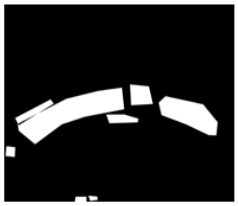
(d) (e) (f)

(g) (h) (i)

Figure 15: Disagreement between polygon-shaped (in blue colour) and rectangle-shaped (in red colour) groundtruth regions. It is found out that evaluation results using the polygon-shaped groundtruth provides a more accurate representation of algorithm's performance.

1.'RESTORAN'
- Polygon-Shaped Ground Truth:
  x:[73,321,343,99]
  y:[692,578,604,736]
- Orientation : Multi-Oriented

2.'TWELVE'
- Polygon-Shaped Ground Truth:
  x:[91,431,653,819,835,561,401,213,99]
  y:[728,574,526,506,638,684,730,852,770]
- Orientation : Curved



3.'12'
- Polygon-Shaped Ground Truth:
  x:[875,1007,1035,883]
  y:[486,498,606,614]
- Orientation : Horizontal

.
.
.

8.'Welcome'
- Polygon-Shaped Ground Truth:
  x:[709,835,935,929,731]
  y:[670,670,696,718,724]
- Orientation : Horizontal

1.'CALIFORNIA'
- Polygon-Shaped Ground Truth:
  x:[306,335,379,424,463,481,460,444,412,372,343,330,313]
  y:[26,28,52,85,104,76,50,58,75,108,102]
- Orientation : Curved

2.'STATE'
- Polygon-Shaped Ground Truth:
  x:[327,361,394,388,371,343,314,302]
  y:[123,174,176,199,203,185,157,125]
- Orientation : Curved

3.'PARKS'
- Polygon-Shaped Ground Truth:
  x:[403,445,461,484,458,412]
  y:[181,153,116,130,185,206]
- Orientation : Curved

.
.
.

14.'*Do Not Care*'
- Polygon-Shaped Ground Truth:
  x:[352,437,434,355]
  y:[137,140,154,159]
- Orientation : -

1.'Clean'
- Polygon-Shaped Ground Truth:
  x:[89,125,249,321,237,177,95]
  y:[490,392,250,306,400,550,538]
- Orientation : Curved

2.'Food'
- Polygon-Shaped Ground Truth:
  x:[285,387,547,563,479,335]
  y:[184,144,106,212,220,284]
- Orientation : Curved

3.'Good'
- Polygon-Shaped Ground Truth:
  x:[615,589,605,717,823,879,787]
  y:[106,180,224,248,334,256,204]
- Orientation : Curved

4.'Taste'
- Polygon-Shaped Ground Truth:
  x:[929,1003,993,937,893,843,899]
  y:[308,490,566,584,446,362,278]
- Orientation : Curved

1.'FISHERMANS'
- Polygon-Shaped Ground Truth:
  x:[555,576,637,745,861,933,912,801,721,667,648,634]
  y:[541,410,323,275,296,329,401,356,382,418,493,566]
- Orientation : Curved

2.'WHARF'
- Polygon-Shaped Ground Truth:
  x:[945,1012,1077,1096,1032,1006,952,925]
  y:[353,415,532,655,637,530,434,419]
- Orientation : Curved

3.'OF'
- Polygon-Shaped Ground Truth:
  x:[634,670,594,562]
  y:[604,716,710,617]
- Orientation : Curved

4.'SAN'
- Polygon-Shaped Ground Truth:
  x:[673,736,790,750,657,622]
  y:[724,785,830,904,823,761]
- Orientation : Curved

5.'FRANSICO'
- Polygon-Shaped Ground Truth:
  x:[799,894,948,1009,1036,1092,1092,1056,936,850,789]
  y:[842,853,835,776,680,697,743,851,941,949,920]
- Orientation : Curved

Figure 16: Examples of Total-Text annotations with polygon-shaped groundtruth. It can be noticed that the polygon-shaped bounding box tightly bounded the text, the annotations are more comprehensive.

1.'Triple-O's'
- Polygon-Shaped Ground Truth:
  x:[585,875,1155,1177,1127,1019,843,823,615]
  y:[531,364,310,422,470,492,578,656,744]
- Orientation : Multi-Oriented

2.'byWhite'
- Polygon-Shaped Ground Truth:
  x:[729,725,757,827,887,885,795,725,677,675,703]
  y:[734,788,890,942,952,1014,994,937,845,758,718]
- Orientation : Curved

3.'SPOT'
- Polygon-Shaped Ground Truth:
  x:[903,987,1011,1029,1085,1045,1003,921]
  y:[948,900,845,784,803,912,970,1014]
- Orientation : Curved

4.'*Do No Care*'
- Polygon-Shaped Ground Truth:
  x:[187,293,367,347,229,245]
  y:[1118,1108,1462,1510,1520,1456]
- Orientation : -

5.'*Do No Care*'
- Polygon-Shaped Ground Truth:
  x:[5,97,83,1]
  y:[1308,1300,1370,1388]
- Orientation : -

1.'SNAPME!'
- Polygon-Shaped Ground Truth:
  x:[2297,2282,2327,2381,2432,2435,2378,2357]
  y:[1376,1250,1106,992,1025,1073,1175,1400]
- Orientation : Curved

2.'Continental'
- Polygon-Shaped Ground Truth:
  x:[2546,2597,2789,2975,2957,2750,2579]
  y:[842,779,770,794,878,854,920]
- Orientation : Curved

3.'TWEETME!'
- Polygon-Shaped Ground Truth:
  x:[3170,3245,3281,3287,3218,3218,3158,3122]
  y:[974,1064,1220,1397,1388,1235,1085,1031]
- Orientation : Curved

4.'Continental'
- Polygon-Shaped Ground Truth:
  x:[2699,2900,3080,3107,2708]
  y:[1763,1751,1739,1826,1850]
- Orientation : Horizontal

5.'TYRETRADERS.COM'
- Polygon-Shaped Ground Truth:
  x:[431,2183,2174,437]
  y:[908,875,1013,1028]
- Orientation : Horizontal

1.'WIGLEY'
- Polygon-Shaped Ground Truth:
  x:[341,429,470,536,636,611,522,442,332]
  y:[102,118,122,129,174,216,186,176,160]
- Orientation : Curved

2.'FIELD'
- Polygon-Shaped Ground Truth:
  x:[658,688,725,802,787,761,646]
  y:[193,209,245,324,356,345,238]
- Orientation : Curved

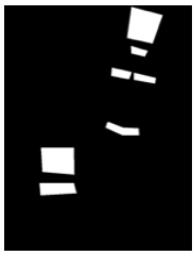
3.'HOME'
- Polygon-Shaped Ground Truth:
  x:[482,607,600,471]
  y:[184,241,276,219]
- Orientation : Curved

.
.
.

12.'PIRATES'
- Polygon-Shaped Ground Truth:
  x:[300,477,468,292]
  y:[265,331,367,303]
- Orientation : Multi-Oriented

1.'17'
- Polygon-Shaped Ground Truth:
  x:[525,655,613,507]
  y:[5,37,147,125]
- Orientation : Multi-Oriented

2.'ONTARIO'
- Polygon-Shaped Ground Truth:
  x:[523,593,581,525]
  y:[155,175,197,185]
- Orientation : Multi-Oriented



3.'TRANS'
- Polygon-Shaped Ground Truth:
  x:[445,527,515,441]
  y:[239,257,285,271]
- Orientation : Multi-Oriented

.
.
.

7.'ONTARIO'
- Polygon-Shaped Ground Truth:
  x:[431,493,555,557,487,417,417,425,433,425]
  y:[449,471,474,500,500,471,476,451,450,451]
- Orientation : Curved

Figure 17: More examples of Total-Text annotations with polygon-shaped groundtruth.