DEEP SALIENCE MAP GUIDED ARBITRARY DIRECTION SCENE TEXT RECOGNITION

Xinhao Liu *Takahito Kawanishi* *Xiaomeng Wu* *Kaoru Hiramatsu* *Kunio Kashino*

NTT Corporation, Kanagawa, Japan 243-0198
{liu.xinhao, kawanishi.takahito, wu.xiaomeng, hiramatsu.kaoru, kashino.kunio} @lab.ntt.co.jp

ABSTRACT

Irregular scene text such as curved, rotated or perspective texts commonly appear in natural scene images due to different camera view points, special design purposes etc. In this work, we propose a text salience map guided model to recognize these arbitrary direction scene texts. We train a deep Fully Convolutional Network (FCN) to calculate the precise salience map for texts. Then we estimate the positions and rotations of the text and utilize this information to guide the generation of CNN sequence features. Finally the sequence is recognized with a Recurrent Neural Network (RNN) model. Experiments on various public datasets show that the proposed approach is robust to different distortions and performs superior or comparable to the state-of-the-art techniques.

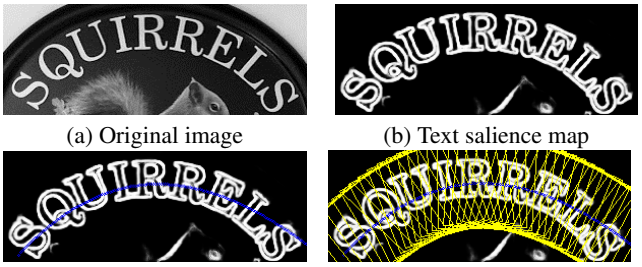
Index Terms— Scene Text Recognition, Fully Convolutional Network, Recurrent Neural Network, Text Salience Map, Arbitrary Direction Text

1. INTRODUCTION AND RELATED WORK

Recognizing the texts from images is a practical task that has attracted increasing attention during recent years. Different from the traditional Optical Character Recognition (OCR) technique, the difficulty of recognizing the natural scene texts comes from the unconstrained conditions. For example, irregular scene texts such as curved, rotated or perspective texts are commonly appearing in scene images due to different camera view points, special design purposes etc. Thus developing a robust text recognition algorithm which can handle different kinds of distortion has become an interesting yet challenging research topic.

As in [1], a complete text detection and recognition system can be divided into various sub tasks such as text localization, word recognition and so on. In this work, we focus on the cropped word recognition task. Many algorithms have been proposed to solve this problem. For character recognition, methods based on features such as Histogram of Gradients (HOG) [2, 3], text specific multi-scale features [4], mid level features [5], low-dimensional attribute [6] or Convolutional Neural Network (CNN) features [7, 8, 9, 10, 11] have been widely explored in existing literature. Among them, the discriminative CNN features have shown its effectiveness and greatly improved the performance. For word recognition, in [10] a lexicon-driven segmentation based method is utilized and by considering the word image as a sequence of characters, in [8] a HMM based and in [12] a Recurrent Neural Network (RNN) based sequence model have been proposed and achieved encouraging results. However, these methods do not have the module to deal with the irregular text such as Fig.1(a) and the performance could also drops if the bounding box is not well labeled.

To recognize the perspective texts, rotation invariant features such as SIFT [13] or circular Fourier-HOG [14] are extracted and utilized to handle the distortion. Improved as they are, the results are still not satisfied. In [15], the distorted images are rectified by a *spatial transform network* (STN) and then recognized with a sequence classification network. However the STN is sensitive to the initialization and the performance could be easily affected if the shape of the actual text is not consistent with the initial fiducial points.



(a) Original image (b) Text salience map

(c) Text centroid curve (d) Guided sliding windows

Fig. 1: Illustration of the proposed method for arbitrary text recognition. Firstly an accurate text salience map (b) is predicted with FCN, and then the text centroid curve (c) is estimated and the sequence of CNN features (d) extracted and recognized with a RNN model.

In real world, the texts are usually curved or rotated and the word images are also not well cropped which decrease the performance of the recognition algorithm. In this work, to address the problem of irregular text, we propose a salience map based approach. In contrast to previous methods, we employ the deep FCN [16] which can learn to compute the text salience map more effectively and efficiently. Then we derive the character centroid curve using the maximum likelihood estimation. Finally the sequence of CNN features is recognized with a RNN model. As the pipeline illustrated in Fig. 1, our approach can explicitly find the positions and rotations of the text and utilize that information to guide the generation of CNN sequence. In the experiments, we show that the proposed approach is superior or comparable with the state-of-the-art for word recognition on various public datasets.

The following of the paper is organized as: the proposed approach is described in Section 2, The implementation and experimental results are described in detail in Section 3. Conclusions and future direction are given in Section 4.

2. PROPOSED APPROACH

2.1. Fully Convolutional Network for Text Salience Map

Fully Convolutional Network (FCN) has been proposed to understand the images from pixel level and has achieved promising results than conventional methods for tasks such as semantic segmentation

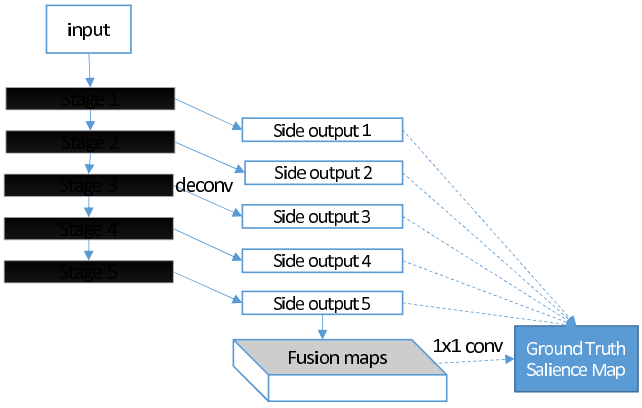
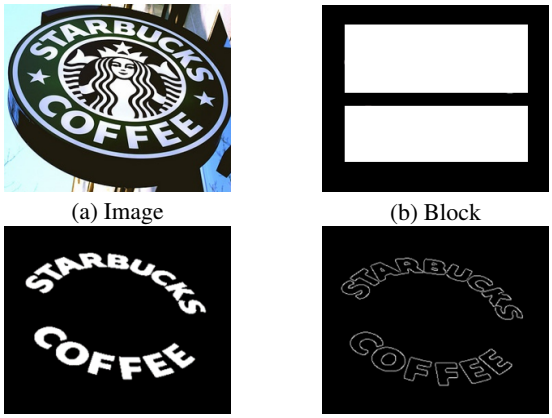


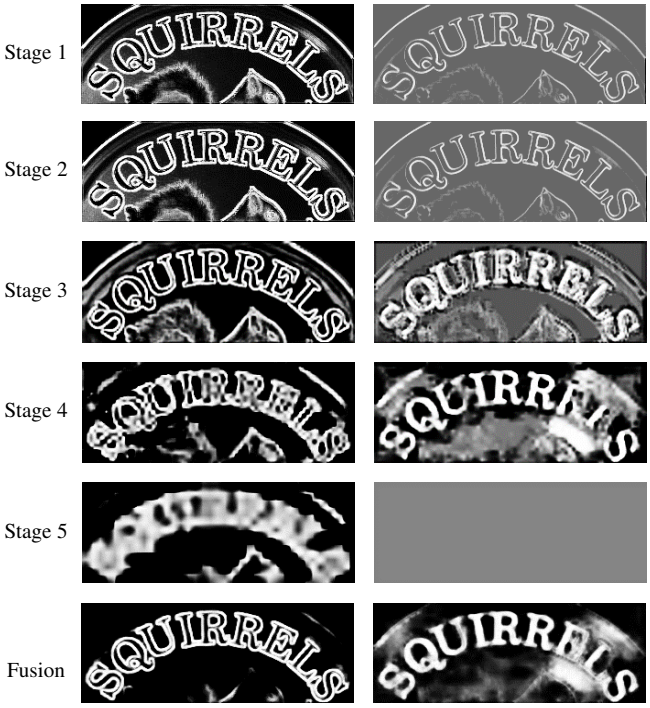
Fig. 2: Network structure for text salience map based on VGG-16 model with deep supervision.



(a) Image (b) Block

(c) Segmentation (d) Edge (Boundary)

Fig. 3: Different types of pixel-level ground truth.



(a) Edge (b) Segmentation

Fig. 4: Output of each stage using different ground truths.

[16], edge detection [17]. We also utilize the FCN to generate the text salience map for its advantages that 1) it can effectively capture the local and global structure of the texts with the convolutional and de-convolutional layers. 2) it is very efficient since all the computation is done within one forward pass. We employ the widely used VGG-16 network [18] with deeply supervised hidden layers [17] as our basic structure. The overall network structure is shown in Fig. 2. There are 5 stages of convolutional layers same with the VGG-16 layer network. The output of each conv layer is feed to the de-convolutional layer [16] to generate a feature map which has the same size with original image. To improve the pixel level labeling accuracy, we utilize the deep supervision strategy proposed in [17] where each side output is also supervised by the ground truth. The side outputs are finally weighted fused through a 1×1 conv layer to get the final prediction. During training, the side output cross entropy losses L_{side} as well as the fused loss L_{fuse} are computed and minimized through stochastic gradient descent method.

There are different types of pixel level ground truth maps. According to the difficulty of annotation, the easiest type is the text block as used in [19]. However, the text block contains very coarse information and cannot capture the accurate shape of the text regions. In this work we consider two other types namely text segmentation and text edge (or boundary) which carry more accurate information. The visual examples of these two types are illustrated in Fig. 3 (c) and (d). Edge map can be easily obtained from the segmentation. To find which is better we conducted experiments with both types and predicted text salience map of two models are shown in Fig. 4. On one hand, we can see that the text salience map of each stage captures the structures at different levels from local to global. Thus we can get the accurate structure of the texts from the final fused map. On the other hand, the fused map trained with the edge map as the ground truth becomes much clearer than by using segmentation since the convolutional operations is more sensitives to the edges.

The spatial size of the feature map shrinks with number of pooling layers, to further reduce the model parameters and avoid the over shrinkage, we also tried another network architecture variant of VGG-16 network by removing the last 2 stages for the model shown Fig. 2. However, in the experiment we find that when the depth is decreased, the quality of generated salience map become relatively poorer with more noise and non-text artifacts.

2.2. Guided Sampling of Sliding Windows

Once the text salience map is computed, we can know the character locations more precisely and sampling the sliding window more accurately with the guidance of the salience map. Pixels with probabilities greater than zero are considered as the foreground pixels and denoted as $D = \{(x_i, y_i)...(x_n, y_n)\}, i = 1, 2, ...n$, where n is the total number of text pixels. We assume the pixels of characters are vertically symmetric. Thus the curve derived from these points should be the curve formed by the characters' centroids. We model the curve of centroids using a k-th order polynomial manner,

$$y_i = f(x_i) = a_0 + a_1 x_i^1 + \cdots + a_k x_i^k, \quad (1)$$

where $a_1, ..., a_k$ are the coefficients need to be estimated. Writing Eq. (1) in matrix form and the maximum likelihood estimation of the coefficients $\mathbf{a}$ is given below,

$$\mathbf{a} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \quad (2)$$

We sample the sliding window along the horizontal axis of the image with a stride of 2 pixels, let x_t denotes the x-axis position of the sliding window, then centroid position of a character is given by

$$(x, y) = (x_t, f(x_t)). \quad (3)$$

The rotation angle θ of the sliding window can be derived from the slope of the curve as,

$$\theta = -\frac{1}{\frac{\partial f}{\partial x}(x_t)}. \quad (4)$$

One example of the estimated curve and sampled sliding window is shown in Fig. 1.

2.3. Recurrent Neural Network for Sequence Recognition

For sequence recognition, the methods based on bi-directional Recurrent Neural Network with LSTM [20] structure and Connectionist Temporal Classification (CTC) [21] have show promising results. To recognize the generated sequence of sliding windows as character sequence, in this work we adopt the same architecture as in [12]. In the network, there are two bi-directional LSTM layers with 256 hidden units. For consistence, we re-use the VGG-16 network and extract the output of last fully connected layer as the feature vector (512 dims) to train the RNN network. Usually a lexicon or word list is provided to correct the prediction to the correct word in the language. Given the raw prediction results $\mathbf{w}$ from RNN, the recognition result $\mathbf{s}^*$ is derived as the word in the lexicon which maximize the probability,

$$\mathbf{s}^* = \arg\max_{\mathbf{s}\in D}(\log \mathrm{P}(\mathbf{s}|\mathbf{w})), \quad (5)$$

where D is the lexicon, and $\mathrm{P}(\mathbf{s}|\mathbf{w})$ is the probability of lexicon word $\mathbf{s}$ given the raw sequence $\mathbf{w}$.

2.4. Hybrid Method for Robust Text Recognition

The proposed method is not without limit. For the low resolution or heavily corrupted images, the computation of text salience map is theoretically impossible. In that situation the recognition performance of simple sliding window approach which is based on the original human annotated bounding box is better. To fix this problem, we propose to use a hybrid method. The final prediction $\mathbf{s}_{final}$ is the result of simple sliding window based prediction s_{simple} and guided sliding window based prediction s_{guided} with higher probability,

$$\mathbf{s}_{final} = \max_{\mathbf{s}}\{\mathrm{P}(\mathbf{s}_{simple}), \mathrm{P}(\mathbf{s}_{guided})\}. \quad (6)$$

3. EXPERIMENTAL RESULTS

In this section we present the experimental setting and recognition results for our algorithm. Firstly we train a VGG-16 network for character classification with the data collected with various datasets in [10], then we initialize the salience map network with this pretrained model. We manually labeled a dataset which contains 18282 word images with binary segmentations from 3000 natural scene images and use it as training data for the text salience map network. For the RNN sequence recognition model, similar as [12] we use the 3600 word images collected from training set of ICDAR03, SVT, KAIST, and IIIT5K dataset. For test, we resize each test image to fixed height of 40 pixels while keeping its original aspect ratio and center crop each sliding window with size 32×32 to compute the CNN features.

3.1. Results on Word Recognition

Firstly we evaluate the proposed algorithm on widely used ICDAR 2003, SVT-WORD and IIIT5K datasets. Words with less than two characters are ignored thus leaving 862 test images for the ICDAR 2003 test set and 647 for the SVT-WORD. The recognizer is tested with a lexicon of different size. The lexicon is formed by the ground truth word plus a number of distractors. For the ICDAR 2003 dataset, the distractors are all words in ICDAR 2003 test set (ICDAR03-full) or 50 randomly selected words (ICDAR03-50) [22] and for SVT-WORD, there are 50 selected words [2]. The word images in IIIT5K datasets are associated with two lexicons of 50 words and 1000 words.

Table 1 shows the word recognition results for different algorithms on ICDAR 2003, SVT-WORD and IIIT5K datasets. We use the simple sliding window based RNN sequence recognition model with VGG16 features as our baseline. The results with guided sliding window and hybrid method are reported. The proposed guided method improves the baseline and exising work [12] on most datasets except SVT-word dataset. The SVT word dataset contains many low-resolution images and in that case the proposed approach has better performance using the hybrid mode since they are difficult to computed the accurate text salience map. For dataset with more irregular texts such as IIIT5K the guided sampling mode itself improves the baseline more than the hybrid mode.

Methods [15, 25] employ very large scale dataset (9×10^6 word images) and achieve high performance. However, in [25] by considering the each word among a 90, 000 sized English lexicon as one class, the model requires large number of parameters and is not flexible to extend to new or unknown words once trained. In [15] the initial fudical points of the rectification should be carefully set according to prior knowledge and the performance could be degraded when the actual text shapes are not consistent with the initialization. On contrast, the proposed approach is more flexible and robust to different distortions. We believe the performance of proposed approach could also gained through large scale dataset training.

3.2. Results on Arbitrary Direction Word Recognition

We also test our approach on more challenging SVT perspective [13] and MSRA-TD500-WORD dataset [26]. The SVT perspective dataset contains 639 word images and the images are intentionally picked from Google Street View with a variety of viewpoints and orientations. MSRA-TD500-WORD is also dataset that includes many texts of arbitrary orientations and perspective texts. The results for these two datasets are shown in Table 2. We can see a large gain in the performance compared with the baseline and other existing algorithms, the improvement indicates that the proposed approach is robust against the distortions and rotations.

For the speed, tested with ICDAR03 dataset with 50 sized lexicon, the proposed guided approach runs at 0.74s per image on aver-

Table 1: Cropped word recognition accuracy (%) on the ICDAR 2003, SVT-WORD and IIIT5K datasets. The numbers in the bracket are lexicon size.

Method	ICDAR03 (50)	ICDAR03 (full)	SVT-WORD (50)	IIIT5K (50)	IIIT5K (1000)
Wang et al.[22]	76.0	62.0	57.0	-	-
Mishar et al. [23]	81.8	67.8	73.2	-	-
Wang & Wu [7]	90.0	84.0	70.0	-	-
Shi et al. [24]	87.4	79.3	73.5	-	-
Yao et al. [4]	88.5	80.3	75.9	80.2	69.3
Almazan et al. [6]	-	-	87.0	88.6	75.6
Alsharif et al. [8]	93.1	88.6	74.3	-	-
Gordo et al. [5]	-	-	90.7	93.3	86.6
Jaderberg et al. [10]	96.2	91.5	86.1	-	-
He et al. [12]	**97.0**	93.8	**93.5**	94.0	91.5
Proposed (Baseline)	96.6	93.4	90.9	95.1	90.8
Proposed (Guided)	96.8	94.2	90.1	**96.4**	**92.3**
Proposed (Hybrid)	**97.0**	**94.4**	91.5	96.1	92.1
	Training with very large additional dataset				
PhotoOCR [9]	-	-	90.4	-	-
Jaderberg et al. [25]	**98.7**	**98.6**	**95.4**	**97.1**	**92.7**
Shi et al. [15]	98.3	96.2	**95.5**	96.2	**93.8**

Table 2: Cropped word recognition accuracy (%) on the SVT-Perspective Word and MSRA-TD500 Word dataset.

Method	SVT-Perspective (50)	SVT-Perspective (Full)	MSRA-TD-500 (Full)
K. Wang et al. [22]	40.5	26.1	44.5
Mishar et al. [23]	45.7	67.8	73.2
T. Wang et al. [7]	40.2	32.4	20.8
Phan et al. [13]	62.3	42.2	58.4
Zhou et al. [14]	67.0	45.7	65.4
Proposed (Baseline)	78.6	62.2	71.7
Proposed (Guided)	**80.9**	63.1	**78.2**
Proposed (Hybrid)	79.7	**64.2**	76.7

age with an Intel Xeon E5-2650 2.6 GHz $\times$ 2 machine with GPU. Fig. 5 shows some correctly recognized examples and failure cases. Images and salience maps are resized to similar scale for better visual purpose. Many of the failures are due to low resolution, fancy font styles and so on. For these images the computation of text salience map is very challenging and some of them are even difficult for humans to read.

4. CONCLUSION

In this paper, a salience map based text recognition approach is proposed to handle the problem of irregular text. The deep FCN [16] is learned and utilized to calculate the text salience map for its high performance and efficiency. Then the positions and rotations of the text are estimated using maximum likelihood and utilized to guide the generation of CNN sequence. Finally the sequence of CNN features is recognized with a Recurrent Neural Network model. As shown in the experiments, the proposed approach is robust to many commonly appeared distortions like curved, rotate, perspective texts and so on. In future work, we will extend the FCN based salience computation method to the text detection task and build an end-to-end recognition framework.

(a)

(b)

Fig. 5: Correctly recognized examples (a) and incorrect ones (b).

5. REFERENCES

[1] Simon M Lucas, Alex Panaretos, Luis Sosa, Anthony Tang, Shirley Wong, Robert Young, Kazuki Ashida, Hiroki Nagai, Masayuki Okamoto, Hiroaki Yamamoto, et al., "ICDAR 2003 robust reading competitions: entries, results, and future directions," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 7, no. 2-3, pp. 105–122, 2005.

[2] Kai Wang and Serge Belongie, "Word spotting in the wild," in *ECCV*, pp. 591–604. Springer, 2010.

[3] Shangxuan Tian, Shijian Lu, Bolan Su, and Chew Lim Tan, "Scene text recognition using Co-occurrence of Histogram of Oriented Gradients," in *International Conference on Document Analysis and Recognition*. IEEE, 2013, pp. 912–916.

[4] Cong Yao, Xiang Bai, Baoguang Shi, and Wenyu Liu, "Strokelets: A learned multi-scale representation for scene text recognition," in *CVPR*. IEEE, 2014, pp. 4042–4049.

[5] Albert Gordo, "Supervised mid-level features for word image representation," in *CVPR*, 2015, pp. 2956–2964.

[6] Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny, "Word spotting and recognition with embedded attributes," *TPAMI*, vol. 36, no. 12, pp. 2552–2566, 2014.

[7] Tao Wang, David J Wu, Andrew Coates, and Andrew Y Ng, "End-to-end text recognition with convolutional neural networks," in *ICPR*. IEEE, 2012, pp. 3304–3308.

[8] Ouais Alsharif and Joelle Pineau, "End-to-end text recognition with hybrid HMM maxout models," in *ICLR*, 2014.

[9] Alessandro Bissacco, Mark Cummins, Yuval Netzer, and Hartmut Neven, "PhotoOCR: Reading text in uncontrolled conditions," in *ICCV*. IEEE, 2013, pp. 785–792.

[10] Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman, "Deep features for text spotting," in *ECCV*, pp. 512–528. Springer, 2014.

[11] Xinhao Liu, Takahito Kawanishi, Xiaomeng Wu, and Kunio Kashino, "Scene text recognition with high performance CNN classifier and efficient word inference," in *ICASSP*. IEEE, 2016.

[12] Pan He, Weilin Huang, Yu Qiao, Change Loy Chen, and Xiaoou Tang, "Reading scene text in deep convolutional sequences," in *AAAI*, 2016.

[13] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan, "Recognizing text with perspective distortion in natural scenes," in *ICCV*, 2013, pp. 569–576.

[14] Yu Zhou, Shuang Liu, Yongzheng Zhang, Yipeng Wang, and Weiyao Lin, "Perspective scene text recognition with feature compression and ranking," in *ACCV*. Springer, 2014, pp. 181–195.

[15] Baoguang Shi, Xinggang Wang, Pengyuan Lv, Cong Yao, and Xiang Bai, "Robust scene text recognition with automatic rectification," in *CVPR*, 2016.

[16] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.

[17] Saining Xie and Zhuowen Tu, "Holistically-nested edge detection," in *ICCV*, 2015, pp. 1395–1403.

[18] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[19] Zheng Zhang, Chengquan Zhang, Wei Shen, Cong Yao, Wenyu Liu, and Xiang Bai, "Multi-oriented text detection with fully convolutional networks," in *CVPR*, June 2016.

[20] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*. ACM, 2006, pp. 369–376.

[22] Kai Wang, Boris Babenko, and Serge Belongie, "End-to-end scene text recognition," in *ICCV*. IEEE, 2011, pp. 1457–1464.

[23] Anand Mishra, Karteek Alahari, and CV Jawahar, "Scene text recognition using higher order language priors," in *BMVC*. BMVA, 2012.

[24] Cunzhao Shi, Chunheng Wang, Baihua Xiao, Yang Zhang, Song Gao, and Zhong Zhang, "Scene text recognition using part-based tree-structured character detection," in *CVPR*. IEEE, 2013, pp. 2961–2968.

[25] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Reading text in the wild with convolutional neural networks," *International Journal of Computer Vision*, vol. 116, no. 1, pp. 1–20, 2016.

[26] Cong Yao, Xiang Bai, Wenyu Liu, Yi Ma, and Zhuowen Tu, "Detecting texts of arbitrary orientations in natural images," in *CVPR*. IEEE, 2012, pp. 1083–1090.