2017 14th IAPR International Conference on Document Analysis and Recognition

End-to-end Scene Text Recognition in Videos Based on Multi Frame Tracking

Xiaobing Wang, Yingying Jiang, Shuli Yang, Xiangyu Zhu, Wei Li, Pei Fu, Hua Wang, Zhenbo Luo
Machine Learning Lab
Samsung R&D Institute of China, Beijing
Beijing, China
{x0106.wang, yy.jiang} @samsung.com

***Abstract*—Text detection and recognition in scene images and videos attract much attention in computer vision recently. However, most existing text detection and recognition methods only focus on static images. In this paper an end-to-end scene text recognition method based on multi frame tracking is proposed for text in videos, in which temporal information is employed to improve performance. First, an end-to-end text recognition method based on a unified deep neural network is used to detect and recognize text in each frame of the input video. Then, multi frame text tracking is employed through associations of texts in current frame and several previous frames to obtain final results. Experiments on ICDAR datasets demonstrate that the proposed method outperforms the state-of-the-art methods in end-to-end video text recognition.**

Keywords—end-to-end text recognition; text in videos; deep neural network; multi frame tracking

I. Introduction

Text information plays an important role in images and videos, which can be widely applied in content based retrieval, assistive navigation, automatic translation, industrial automation, scene understanding and so on. An end-to-end text information recognition system always includes two steps [1]: text detection, in which text regions are labeled out using bounding boxes, and text recognition, in which text information is retrieved from the detected text regions using optical character recognition (OCR) or other technologies. Therefore, text detection and recognition attract much attention these years. Although many scene text detection and recognition methods have been proposed [2, 3, 4, 5, 6, 7], most of them only focus on static images and do not work well for videos. However, videos are also common in our life and video text detection and recognition are focused on in this paper.

Scene text detection and recognition in videos are challenging problems in computer vision. Except the challenges existing in static scene text images such as complex backgrounds, uneven illumination, the variety of text, and so on, videos present some different challenges. The quality of the image in video is generally worse than static images, due to motion blur and out of focus issues. Meanwhile, blocking artifacts maybe created by video compression. Besides, to employ the temporal information in videos, text tracking is needed while the presence of occlusions may affect it. Therefore, accurate text detection, recognition and tracking are all needed for robust text reading from videos.

End-to-end text recognition of each frame is an important step for end-to-end text recognition in videos. Although many methods have been proposed for text detection and recognition, Convolutional Neural Network (CNN) based methods [3, 6, 7] become popular and show better performances than other methods. Therefore, to accurate detect and recognize text in each frame of a video sequence, an end-to-end scene text recognition method based on a unified deep neural network is proposed in this paper. Meanwhile, text tracking is essential for text detection and recognition in videos, in which the methods of object tracking is used. Because text detection and recognition of each frame are finished before this step, a multi frame tracking method based on the detection and recognition results is used in this paper, in which each frame is associated with several previous frames for accurate tracking.

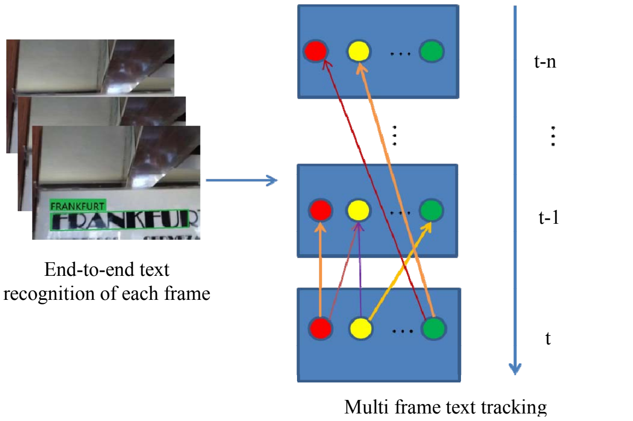
FRANKFURT
FRANKFURT
t-n
t-1
t
End-to-end text recognition of each frame
Multi frame text tracking

Fig. 1. The framework of the proposed end-to-end video text recognition method.

An end-to-end scene text recognition method based on multi frame tracking is proposed for detecting and recognizing text in videos in this paper. As shown in Fig.1, two steps are included in the proposed method: end-to-end text recognition of each frame and multi frame text tracking. When a video sequence is given, an end-to-end text recognition method based on a unified deep neural network is applied to each frame of the video. Then, based on the detection and recognition results of each frame, multi frame text tracking is done through associations of these results in current frame and several previous frames. Now accurate text tracking trajectory, final text detection and recognition results are simultaneously obtained. The proposed method is tested on ICDAR datasets and experimental results show that its performance is much better than the state-of-art methods in end-to-end video text recognition.

The rest of this paper is organized as follows. Section II introduces the related work about text detection, recognition and tracking. Section III describes the end-to-end text

2379-2140/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDAR.2017.207

recognition of each frame in the input video sequence. The multi frame text tracking method based on both text detection and recognition results is presented in Section IV. Section V demonstrates the benchmarks and experimental results to validate the proposed method. Finally, conclusions are stated in Section VI.

II. Related Work

Scene text detection and recognition have been the research foci in computer vision for several years and many methods have been proposed to solve them [2, 3, 4, 5, 6, 7]. Before the appearance of deep learning methods [8], these methods can be divided into two groups: region-based and connected component(CC)-based. In region-based methods [2] local image regions (always sliding windows) are used to detect and recognize text, while in CC-based methods [4, 5] an image is segmented into CCs first and then text is detected and recognized from these CCs. However, with the rise of deep learning methods, CNN based methods become popular in object detection and recognition [9, 10]. As text is a special kind of object, CNN based methods can also be applied to scene text detection and recognition [3, 6, 7], with them much better performances are achieved. Therefore, CNN based methods are considered by us. For object detection, the state-of-the-art pipeline is Faster R-CNN [11]. While for text recognition CNN and Long-Short Term Memory (LSTM) based methods [7] show better performance. Because CNN features are needed in both text detection and recognition, the convolution layers of the two steps are shared in the this paper. Then, a unified deep neural network based on CNN and LSTM is proposed.

Besides, text tracking is also an import step for text detection and recognition in videos. The existing text tracking methods can be divided into three groups [12]: template matching, Bayesian framework and tracking-by-detection based methods. Template matching based methods [13] seek the most similar region as the template image. Bayesian framework based methods usually use particle filter [14] and Kalman filter [15] to track text. Moreover, the tracking-by-detection methods [16,17,18] track text through the association of detected results in successive frames. In the proposed method a tracking-by-detection method is employed. Meanwhile, not only successive frames, the text detection and recognition results between current frame and several previous frames are all associated for more accurate tracking, as shown in Fig.1.

III. End-to-end Text Recognition of Each Frame

Given a video sequence, the first step of the proposed method is end-to-end text recognition of each frame in it. Although many methods have been proposed for solving this problem, CNN based method become popular [3, 6, 7] with the rise of deep learning [8] and show much better performances than previous methods. Therefore, a CNN based method is employed here. Its architecture is shown in Fig.2.

The architecture of the network used here is constructed based on the VGG16 Faster R-CNN [11]. A similar region proposal network (RPN) is used for text proposal with anchor sizes are $\{4, 8, 16\}$ here. Moreover, the proposed network has the same convolution layers, Region of Interest (ROI) pooling layer and fully connected layers as it, except the output bounding box for a detected text region is defined as $(x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4)$ here, as shown in Fig.3. That is because text may have different arrangement orientations and non-horizontal text such as slant text may exist in real scenes. Using the coordinates of the four verticals of the output bounding box to represent a text region is more accurate.

When a text region is detected, it is possible not horizontal. So that the text region is needed to be rotated to horizontal orientation and then using sliding windows to obtain the ROIs for recognition. However, such operations are complex. Instead of it, sliding windows along the medial axis of the detected text region are used to obtain ROIs. Based on the features of them from the convolution feature map and LSTM layer, text in detected regions can be recognized.

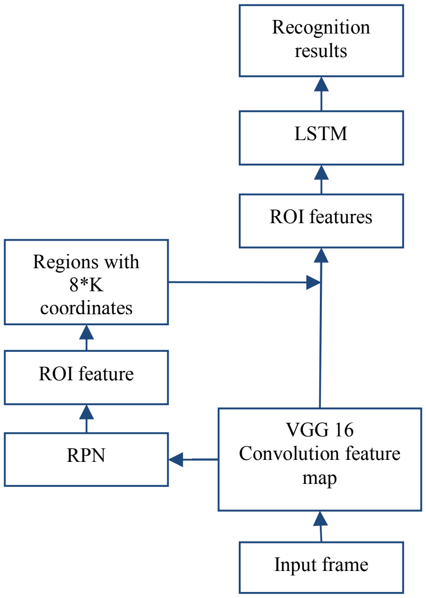

Fig. 2. The architecture of the unified deep neural network for end-to-end text recognition of each frame in a video sequence. In the proposed method covolution layers, ROI pooling layer and fully connected layers are shared for text detection and recognition.

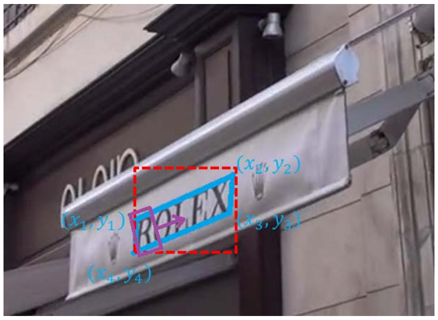

Fig. 3. The representation of a bounding box. The blue bounding box obtained in this way is more accurate than the red bounding box. With the sliding windows along the medial axis of the text region, ROIs for text recogintion are obtained.

In the proposed end-to-end text recognition method, the convolution feature map is shared for detection and recognition. Therefore, the computation of convolution layers only needs one time by doing these operations on the input frame.

IV. Multi Frame Text Tracking

Compared with static images, videos contain temporal information, which can be employed to improve text detection and recognition performances. Therefore, to use the temporal information, the second step of the proposed method is text tracking through which the positions of same text in different frames are determined. Here a tracking-by-detection method is considered by us, which is implemented through the association of the text detection and recognition results in successive frames. However, some text regions may be missed in some frames as shown in Fig.4. If only association of successive frames is considered, the tracking trajectory may be broken when text in one of them is missed. Therefore, multi frame text tracking is used here, in which associations between current frame and several previous frames are all considered.

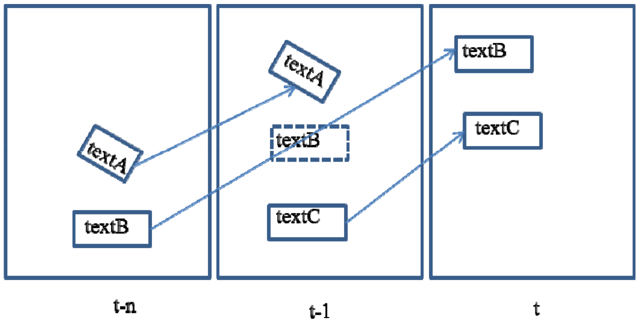

Fig.4. Multi frame text tracking with associations between current frame and several previous frames. Text regions with solid bounding boxes are detected while the one with dotted bounding box is missed.

In text tracking of the proposed method, association of the current frame f_t and one previous frame f_{t-i} is the key problem, which is implemented by computing the similarities or differences between the text regions in them. For two text regions a,b with region a in current frame f_t and region b in frame f_{t-i}, if they are the same region, their spatial positions may be close and overlaps between them may exist. To measure the relations, the scores related to the spatial distance of two text regions $S_{dis}^{spatial}$ and their overlap ratio S_{op}^{area} are computed as follows:

$$S_{dis}^{spatial} = \frac{mean(|x_i^a - x_i^b|)}{\max(W_a, W_b)} + \frac{mean(|y_i^a - y_i^b|)}{\max(H_a, H_b)} \quad (1)$$

$$S_{op}^{area} = \frac{Area(a \cap b)}{Area(a)} \quad (2)$$

Here, (x_i^a, y_i^a) and $(x_i^b, y_i^b), i \in \{1,2,3,4\}$ are the coordinates of the four vertexes of the two regions, H_a, H_b, W_a, W_b are their heights and widths and $mean()$ represents the average value.

Meanwhile, text in detected text regions are recognized as text strings. Though the same text in different frames may be recognized as different text strings, the edit distance should be small. Therefore, a score related to the edit distance between the recognition results of two text regions S_{dis}^{edit} is also needed. It is defined as:

$$S_{dis}^{edit} = \frac{2 * EditDis(str_a, str_b)}{len(str_a) + len(str_b)} \quad (3)$$

Here, str_a and str_b are the text recognition results of the two regions.

Moreover, a region in current frame is more possible to match another region in nearby frames. The offset between the frame ids of the two regions should be considered and the score related to it S_{offset}^{frame} is defined as:

$$S_{offset}^{frame} = abs(id_a - id_b) \quad (4)$$

Based on the scores computed above, the difference between the two regions can be defined as:

$$S_{ab} = k_1 * S_{dis}^{spatial} + k_2 * S_{op}^{area} + k_3 * S_{dis}^{edit} + k_4 * S_{offset}^{frame} \quad (5)$$

where k_1, k_2, k_3, k_4 are the weights of the four scores and they are set as 1, 1, 10, 0.2 in the proposed method. If $S_{ab} < S_T$, S_T is the match threshold and is set as 8 in here, the two regions are similar and region a match with region b .

A region in frame f_t may match with several regions in frame f_{t-i} and vice versa. To find optimal matches for regions in the two frames, Hungarian algorithm [19] is used based on the difference scores of these matched regions. Now each region in current frame may find an optimal marched region in a previous frame.

When the frames of a video sequence is processed with end-to-end text recognition, text tracking is implemented through the associations between each frame and its previous n frames and $n = 5$ is used here. The process of multi frame text tracking is described as follows:

- *For $i = 1 : num$, num is the number of frames of the input video sequence, chose frame i as current frame. Chose text regions in frame i , which do not match with any region in previous frames and generate a temp current frame.*
- *For $j = i - 1 : i - n$,chose frame j as the previous frame. Chose text regions in j which do not match with any region in later frames and generate a temp previous frame.*
- *Compute the differences between the regions in the temp current frame and the temp previous frame and obtain the optimal matches though Hungarian algorithm. Label the matched text regions in frame i and j .*

- *End the process until all frames in the video sequence are processed and the tracking trajectory is obtained.*

When the tracking trajectory is obtained, the detection and recognition results can be modified. The regions appeared less than 5 times are considered as false alarms and removed. Meanwhile, missed regions can be recovered through the positions of matched regions in the trajectory. Moreover, the recognition results are improved by using the text strings appeared most frequently in the trajectory as the final recognition results. Now the final end-to-end text recognition results of the input video sequence is obtained.

V. Experiments

A. Datasets and Evaluation Criteria

To evaluate the proposed method for end-to-end text recognition in videos, it is evaluated on ICDAR 2015 video text dataset [20], which is the benchmark for video text detection and recognition. The dataset consists of a training set of 25 videos (13,450 frames) and a test set of 24 video (14,374 frames) . It is collected in different countries to include text in different languages. The video sequences correspond to 7 high level tasks, that represent typical real-life applications and cover indoors and outdoors scenarios. Moreover, 4 different cameras for different sequences are used to cover a variety of possible hardware used.

For video text detection and recognition, text tracking is essential. So that, the evaluation is based on an adaptation of the CLEAR-MOT framework [21] for multiple object tracking. Three criteria are provided for each method, including: Multiple object tracking precision (MOTP), multiple object tracking accuracy (MOTA) and average tracking accuracy (ATA). MOTP is used to show the ability of the tracker to estimate precise object positions and MOTA accounts for all object configuration errors made by the tracker. Besides, ATA provides an object tracking measure that penalizes fragments both in temporal and spatial dimensions. For video text recognition, a vocabulary for a video is provided, which is strongly contextualized. Because only the results with strongly contextualized text recognition are listed in ICDAR 2015 Robust Reading Competition (RRC) Challenge 3: Reading Text in Videos, the proposed method is evaluated with strong vocabularies in this paper.

In the proposed method, end-to-end scene text recognition of each frame is an important step. To evaluate the performance of this step, ICDAR 2013 scene image dataset [22] is used here, which is a benchmark for scene text detection and recognition evaluation. The dataset consists of 462 images containing text in a variety of colors and fonts on many different backgrounds and in various orientations. Among of them, 229 images are used for training and 233 images are used for testing. These images are captured from natural scenes with a digital camera using auto focus and natural lighting.

For ICDAR 2013 scene image dataset, three criteria including recall, precision and f score [22] are used to evaluate the end-to-end text recognition results. Recall represents the ratio of the number of correctly recognized words to the total number of words in the dataset while precision represents the ratio of the number of correctly recognized words to the total number of recognized words. F score is single measure of quality by combining recall and precision. Because end-to-end text recognition results in videos are evaluated with strong vocabularies, text recognition results of scene images are also evaluated in the same way.

Because the two datasets mentioned above were used in ICDAR 2015 RRC, the results on them can be evaluated through ICDAR 2015 RRC Platform to obtain the performances.

B. Experiments on Scene Images

The proposed method is first evaluated on the ICDAR 2013 scene image dataset to show the performance of end-to-end text recognition of each frame step. Because text detection is an important step for end-to-end text recognition, the detection results and end-to-end results are both shown here.

To train the network for end-to-end text recognition, it is initialed with VGG16 Faster R-CNN model [11]. Then the layers for text detection are trained with a training dataset of 50,000 images, including the training set of ICDAR 2013 scene image dataset, the training set of ICDAR 2015 video text dataset and images collected by us. Now, the layers for text recognition are initialed with the layers for text detection and then trained with our 200k synthetic images using the method of Gupta et.al [23] to obtain the final network for end-to-end text recognition.

TABLE I. Text Detection Results on ICDAR 2013 Scene Image Dataset

Method	Precision (%)	Recall (%)	F score (%)
The proposed	90.50	**80.18**	**85.03**
CannyText [24]	86.26	78.45	82.17
CTPN [25]	92.77	73.72	82.15
Text-CNN [26]	**92.79**	72.89	81.65
RRPN [27]	90.22	71.89	80.02
FASText [28]	84.00	69.30	76.80
USTB_TexStar [4]	88.47	66.45	75.89
TextSpotter [5]	87.51	64.84	74.49

TABLE II. Edn-to-end Text Recognition Results with strong vocabularies on ICDAR 2013 Scene Image Dataset

Method	Precision (%)	Recall (%)	F score (%)
The proposed	**93.17**	81.79	**87.11**
VGGMaxBBNet [6]	89.63	**82.99**	86.18
TextProposals [29]	-	-	81.16
Stradvision-1	88.66	75.03	81.28
TextSpotter [5]	85.91	69.79	77.02

Table I shows the text detection results on ICDAR 2013 scene image dataset. Among of them, USTB_TexStar and

TextSpotter are the leading participants in text detection task of ICDAR 2015 RRC Challenge 2: Focused Scene Text [20]. The proposed method achieves the highest recall (80.18%) and the highest f score (85.03%) on this dataset. Though the Text-CNN method has the highest precision (92.79%), its recall is much lower and its f score is lower than the proposed method. The highest f score means that the proposed method has better performance than other methods in this table. The proposed method is effective and achieves state-of-art for scene text detection.

Table II shows the end-to-end text recognition results with strong vocabularies of the proposed method and several state-of-art methods on ICDAR 2013 scene image dataset. Among of them, VGGMaxBBNet and Stradvision-1 are the leading participants in end-to-end text recognition task of ICDAR 2015 RRC Challenge 2 [20]. And TextSpotter is the baseline in the competition. The proposed method achieves the highest precision and the highest f score on this dataset. Moreover, its recall only a little lower than VGGMaxBBNet, but precision is obviously higher. So that, the proposed method has better performance than the other methods of this table on this dataset. Therefore, end-to-end text recognition step of the proposed method also achieves state-of-art for this task.

C. Experiments on Videos

To evaluate the performances of the proposed method for text in videos, it is tested on ICDAR 2015 video text dataset. Here, text are not only detected and recognized in each frame, but also tracked over the video sequence.

TABLE III. TEXT DETECTION RESULTS ON ICDAR 2015 VIDEO TEXT DATASET

Method	MOTP (%)	MOTA (%)	ATA (%)
The proposed	69.74	**57.45**	**55.74**
Deep2Text I (Video)	71.01	40.77	45.18
USTB-TexVideo	71.33	49.33	41.31
AJOU [30]	**73.25**	53.45	38.77

Table III shows the text detection results of the proposed method and several state-of-art methods on this dataset. Here, Deep2Text I (Video), USTB-TexVideo and AJOU are the leading participants in text detection task of ICDAR 2015 RRC Challenge 3 [20]. MOTPs and MOTAs show the performances of text detection steps in these methods. The proposed method shows better performance in text detection than other methods with the highest MOTA (57.45%) and a little lower MOTP (69.74%). Meanwhile, it has the highest ATA (55.74%) which is higher than the second highest ATA by 10%, that means the proposed method has superior tracking performance and multi frame tracking in it works well . The propose method has better performance than other methods in this table for text detection in videos.

TABLE IV. EDN-TO-END TEXT RECOGNITION RESULTS ON ICDAR 2015 VIDEO TEXT DATASET

Method	MOTP (%)	MOTA (%)	ATA (%)
The proposed	**70.15**	**68.63**	**60.25**
TextSpotter [5]	69.51	59.83	41.84
Stradivision-1	69.21	56.54	28.53
USTB-TexVideo	65.08	45.82	19.85

Table IV shows the end-to-end text recognition results of the proposed method and several state-of-art methods on this dataset. Here, TextSpotter is the baseline in end-to-end text recogniton task of ICDAR 2015 RRC Challenge 3 [20], while Stradivision-1 and USTB-TexVideo are the leading participants in this task. The proposed method shows better performance in end-to-end text recognition than other methods with the highest MOTP (70.15%) and the highest MOTA (68.63%). Especially, the MOTA is about 9% higher than the second highest MOTA. That means end-to-end text recognition of the propose method outperforms the other methods in the table. Meanwhile, it has the highest ATA (60.25%) which is about 20% higher than the second highest ATA, that means the proposed method has superior tracking performance and achieves much better performance for end-to-end text recognition in videos. Text in videos can be accurately detected, recognized and tracked with the proposed method as shown in Fig.5.

Though the proposed method achieves state-of-art in end-to-end video text recognition, its performance is still low, that means this task for videos is still very challenging.

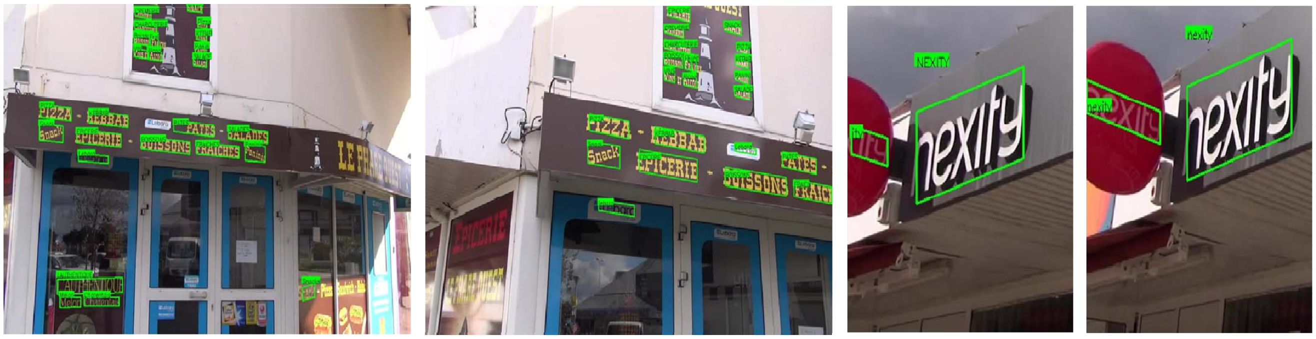

Fig.5. End-to-end text recognition results of the proposed method on sample videos. The left two frames belong to the same video seqence while the right two ones are in another video sequence.

VI. Conclusions

In this paper, a multi frame tracking based method is proposed for end-to-end scene text recognition in videos. It includes two steps: end-to-end text recognition of each frame, in which text in each frame of a video sequence are detected and recognized, and multi frame text tracking, in which recognized text are tracking over the video sequence and final text recognition results are obtained. Experiments on ICDAR datasets show that the proposed method outperforms state-of-art methods in end-to-end video text recognition.

Though the proposed method achieves better performance, it is still low for commercial applications with ATA only 60.25%. That means text detection, recognition and tracking in videos are still challenging and much work is needed in the future.

References

[1] X. Yin, Z. Zuo, S. Tian, and C. Liu, “Text Detection, Tracking and Recognition in Video: A Comprehensive Survey,” IEEE Transactions on Image Processing, vol. 25, no. 6, pp. 2752–2773, 2016.

[2] X. Chen and A. L. Yuille, “Detecting and reading text in natural scenes,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2004, pp. 366–373.

[3] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng, “End-to-End Text Recognition with Convolutional Neural Networks,” Proceedings of the 21st International Conference on Pattern Recognition (ICPR), 2012, pp. 3304–3308.

[4] X. Yin, X. Yin, K. Huang, and H. Hao, “Robust Text Detection in Natural Scene Images,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 5, pp. 970–983, 2014.

[5] L. Neumann and J. Matas, “Real-Time Lexicon-Free Scene Text Localization and Recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 9, pp. 523–527, 2016.

[6] M. Jaderberg, K. Simonyan, A. Vedaldi, and A, Zisserman, “Reading Text in the Wild with Convolutional Neural Networks,” International Journal of Computer Vision, vol. 116, no.1, pp. 1–20, 2016.

[7] B. Shi, X. Bai, and C. Y, “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, vol. PP, no. 9, pp. 1–1, 2016.

[8] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, pp. 436–444, 2015.

[9] D. C. Ciresan, U. Meier, J. Masci, and J. Schmidhuber, “Multi-Column Deep Neural Network for Traffic Sign Classification,” Neural Networks, vol. 32, pp. 333–338, 2012.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” Advances in Neural Information Processing Systems (NIPS), 2012.

[11] S. Ren, K. He, R. Girshick and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” Advances in Neural Information Processing Systems (NIPS), 2015.

[12] S. Tian, W. Pei, Z. Zuo, and X. Yin, “Scene Text Detection in Video by Learning Locally and Globally,” Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI), 2016, pp. 2647–2653.

[13] T. Yusufu, Y. Wang, and X. Fang, “A video text detection and tracking system,” Proceedings of the 2013 IEEE International Symposium on Multimedia (ISM), 2013, pp. 522–529.

[14] C. M. Gracia and M. Mirmehdi, “A framework towards realtime detection and tracking of text,” Proceedings of the Second International Workshop on Camera-Based Document Analysis and Recognition (CBDAR), 2007, pp. 10–17.

[15] C. M. Gracia and M. Mirmehdi, “Real-time text tracking in natural scenes,” IET Computer Vision, vol. 8, no. 6, pp. 670–681, 2014.

[16] X. Liu and W. Wang, “Robustly extracting captions in videos based on stroke-like edges and spatio-temporal analysis,” IEEE Transactions on Multimedia, vol. 14, no. 2, pp. 482–489, 2012.

[17] X. Rong, C. Yi, X. Yang and Y. Tian, “Scene text recognition in multiple frames based on text tracking,” IEEE International Conference on Multimedia and Expo, 2014, pp. 1–6.

[18] S. Tian, X. Yin, Y. Su and H. Hao, “A Unified Framework for Tracking Based Text Detection and Recognition from Web Videos,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PP, pp. 1–1, 2017.

[19] H. W. Kuhn, “The Hungarian Method for the assignment problem,” Naval Research Logistics Quarterly, vol. 2, pp. 83–97, 1955.

[20] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar; S. J. Lu, F. Shafait, S. Uchida, and E. Valveny, “ICDAR 2015 Competition on Robust Reading,” 13th International Conference on Document Analysis and Recognition (ICDAR), 2015, pp. 1156–1160.

[21] K. Bernardin and R. Stiefelhagen, “Evaluating multiple object tracking performance: The CLEAR MOT metrics,” EURASIP Journal on Image and Video Processing, vol. 2008, 2008.

[22] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. Gomez i Bigorda, S. Robles Mestre, J. Mas, D. Fernandez Mota, J. Almaz`an Almaz`an, and L. P. de las Heras, “ICDAR 2013 robust reading competition,” 12th International Conference on Document Analysis and Recognition (ICDAR), 2013, pp. 1115–1124.

[23] A. Gupta, A. Vedaldi, and A. Zisserman, “Synthetic Data for Text Localisation in Natural Images,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2315–2324.

[24] H. Cho, M. Sung, and B. J, “Canny Text Detector: Fast and Robust Scene Text Localization Algorithm,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3566–3573.

[25] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, “Detecting Text in Natural Image with Connectionist Text Proposal Network,” European Conference on Computer Vision (ECCV) , 2016, pp. 56–72.

[26] T. He, W. Huang, Y. Qiao, and J. Yao, “Text-Attentional Convolutional Neural Network for Scene Text Detection,” IEEE Transactions on Image Processing, vol. 25, no. 6, pp. 2529–2541, 2016.

[27] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue, “Arbitrary-Oriented Scene Text Detection via Rotation Proposals” arXiv:1703.01086, 2017.

[28] M. Buta, L. Neumann, and J. Matas, “FASText: Efficient Unconstrained Scene Text Detector,” IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1206–1214.

[29] L. Gomez and D. Karatzas, “TextProposals: a Text-specific Selective Search Algorithm for Word Spotting in the Wild,” Pattern Recognition, in press.

[30] H. I. Koo and D. H. Kim, “Scene text detection via connected component clustering and nontext filtering,” IEEE Transactions on Image Processing, vol. 22, no. 6, pp. 2296–2305, 2013.