Scene Text Detection in Video by Learning Locally and Globally

Shu Tian †, Wei-Yi Pei †, Ze-Yu Zuo, and Xu-Cheng Yin *
Department of Computer Science, School of Computer and Communication Engineering,
University of Science and Technology Beijing, Beijing 100083, China

Abstract

There are a variety of grand challenges for text extraction in scene videos by robots and users, e.g., heterogeneous background, varied text, nonuniform illumination, arbitrary motion and poor contrast. Most previous video text detection methods are investigated with local information, i.e., within individual frames, with limited performance. In this paper, we propose a unified tracking based text detection system by learning locally and globally, which uniformly integrates detection, tracking, recognition and their interactions. In this system, scene text is first detected locally in individual frames. Second, an optimal tracking trajectory is learned and linked globally with all detection, recognition and prediction information by dynamic programming. With the tracking trajectory, final detection and tracking results are simultaneously and immediately obtained. Moreover, our proposed techniques are extensively evaluated on several public scene video text databases, and are much better than the state-of-the-art methods.

1 Introduction

Signage-text is widely used as visual indicators for navigation and notification, and effective text detection and recognition in scene videos is one key factor for various practical applications with reading in the wild, such as assistance for the visually impaired people [Iwatsuka *et al.*, 2004; Chen and Yuille, 2004; Ezaki *et al.*, 2004; Shiratori *et al.*, 2006; Tanaka and Goto, 2007; 2008; Goto and Tanaka, 2009; Sanketi *et al.*, 2011], real-time translator [Haritaoglu, 2001; Shi and Xu, 2005; Petter *et al.*, 2011; Fragoso *et al.*, 2011], robot and user navigation [Aoki *et al.*, 1999; Minetto *et al.*, 2011], and driving assistant system [Wu *et al.*, 2004; 2005].

There are a considerable number of approaches for scene text detection in video [Yin *et al.*, 2016], most of which focus on extracting text with local information, i.e., within individual frames [Petter *et al.*, 2011; Shivakumara *et al.*, 2012; 2013]. At the same time, there are also a few techniques with spatial and temporal information utilization, i.e., detecting text within multiple frames [Fragoso *et al.*, 2011; Minetto *et al.*, 2011; Gomez and Karatzas, 2014], however, without uniformly integrating detection, tracking and their interactions. As we have known, there are a variety of grand challenges for text extraction in scene videos, e.g., heterogeneous background, varied text, nonuniform illumination, arbitrary motion and poor contrast [Ye and Doermann, 2015; Yin *et al.*, 2016]. Consequently, most previous video text detection and recognition methods have limited performance. For example, the recent ATA (Average Tracking Accuracy) result on ICDAR'13 scene video dataset is less than 0.15 [Nguyen *et al.*, 2014].

In this paper, we propose a novel tracking based text detection system in scene videos, which combines detection, tracking and recognition uniformly by learning locally and globally in a unified integration framework. In our system, robust text detection is first performed locally in individual frames. Multi-strategy tracking techniques, e.g., tracking-by-detection, spatial-temporal context learning and template matching, are then used to predict the candidate text position in consecutive frames. Next, tracking trajectories are linked with all detection, recognition and prediction information with a tracking network (graph), where vertex weights are derived from both detection and recognition confidences, and edge weights are based the similarities between the current text block and the predicted ones. Thereafter, an optimal trajectory is learned globally in this network with a dynamic programming algorithm. With this trajectory, final detection and tracking results are simultaneously and immediately obtained. Moreover, our proposed system is verified on a variety of public scene text video databases, i.e., the Minetto [Minetto *et al.*, 2011] and ICDAR'15 datasets [Karatzas *et al.*, 2015a]. Experimental results show that our approach significantly outperforms the state-of-the-art methods on all datasets.

2 Related Work

Text tracking is to determine the position and the time of text continuously and accurately in dynamic video frames. Obviously, text tracking is useful and important for video text detection and recognition. Nowadays, the large number published text tracking can be mainly divided into three main

*Corresponding author: xuchengyin@ustb.edu.cn; † Shu Tian and Wei-Yi Pei contributed equally to this work.

categories: template matching, Bayesian framework and tracking-by-detection based approaches. Template matching based methods implement tracking by seeking the most similar region in the image compared with the template image. For example, Li proposed a gradually improved tracking strategy. The Sum of Square Different (SSD) based image matching is used to track text for a pure translation model [Li and Doermann, 1998]. Moreover, to track text with complex motions, text contour information is utilized to refine the position [Li *et al.*, 2000]. Na and Wen [Na and Wen, 2010] proposed a text tracking scheme based on Scale Invariant Feature Transform (SIFT) features and geometric constraint. Yusufu et al. [Yusufu *et al.*, 2013] used Speeded Up Robust Feature (SURF) and a fast approximate nearest-neighbour search algorithm to track static and rigid moving text objects.

In Bayesian framework based methods, particle filter and Kalman filter are mainly used to track text. Mirmehdi et al. [Merino-Gracia and Mirmehdi, 2007; Merino-Gracia *et al.*, 2012] proposed a near real-time probabilistic tracking framework based on particle filter, where SIFT matching is used to reduce the search space of the particles and identify text regions from one frame to the next. Afterward instead of particle filter, Merino-Gracia and Mirmehdi [Merino-Gracia and Mirmehdi, 2014] presented a complete end-to-end scene text reading system where the unscented Kalman filter (UKF) [Wan and Van Der Merwe, 2000] is utilized to track scene text. Minetto et al. [Minetto *et al.*, 2011] introduced SnooperTrack for automatic detecting and tracking scene text, where the tracking algorithm is based on particle filter. Tanaka and Goto [Tanaka and Goto, 2008; Goto and Tanaka, 2009] developed a wearable camera system for the blind in which the tracking method is also based on particle filter.

The tracking-by-detection method is the association of detected results in successive frames to create the appearances of objects. Thus, the accuracy of text tracking depends largely on the accuracy of text detection. Wang and Wei [Wang and Wei, 2010] proposed a method to track text where Harris corner feature is first extracted, and then Hausdorff distance was taken to measure the dissimilarity. Liu and Wang [Liu and Wang, 2012] proposed a robust extracting captions method in videos based on stroke-like edges and spatio-temporal analysis. These methods are used to track captions rather than scene text which is more challenging with different size, color, contrast, background, orientation and distortion.

Tracking based text detection methods in video are used to reduce alarm and improve the accuracy of detection. The temporal and spatial information generally including the duration of the text, the interval of the starting frame and the ending frame of the same text is often utilized to reduce false alarms [Shiratori *et al.*, 2006; Tanaka and Goto, 2008; Goto and Tanaka, 2009]. Techniques of multiple-frame integration (MFI) techniques, e.g., multi-frame averaging [Wang *et al.*, 2013] and time-based minimum/maximum pixel value searching [Mi *et al.*, 2005], are employed to reduce the influence of the complex background and thereby improve the accuracy of text detection. Furthermore, a method that merge the detection text region with those previously tracked outputs was used to enable false position suppression [Gomez and Karatzas, 2014; Minetto *et al.*, 2011].

Generally speaking, only detecting text locally in individual video frames cannot achieve a high accuracy because of additional arbitrary motion, multi-orientation , multi-scale, poor contrast and degraded text quality challenges. The combination of scene text detection and tracking techniques is an effective way to improve the accuracy of text detection in scene videos. As a result, in this paper, we propose a unified integration framework for tracking based text detection with dynamic programming, which can globally combine a variety of detection and tracking information.

3 Tracking Based Text Detection

3.1 Unified Integration Framework

The unified integration framework is shown in Figure1, the pipeline of which includes three major components. First, in individual frames scene text detection and recognition are conducted and possible text candidates are localized and extracted extensively and locally. Second, in consecutive frames, a variety of tracking approaches (e.g., tracking-by-detection, spatial-temporal context learning and template matching) are performed to predict the candidate text position in the next frame. Finally, all detection and tracking results in consecutive frames are combined with dynamic programming to obtain the preferred results in an optimal tracking trajectory by globally and uniformly integrating detection, tracking, recognition and their interactions. Here, a tracking network (a weighted graph) is first composed of detected and recognized text candidates (vertices/nodes), predicted ones (vertices/nodes), and tracking trajectories (edges). In this graph, vertex weights are derived from both detection and recognition confidences, and edge weights are computed with the similarities between the current text block and its predicted ones. A dynamic programming algorithm is then performed on the weighted graph, and an optimal tracking trajectory is learned globally (see the RED trajectory as an example in Figure1). Afterwards, detection and tracking results in each frame are easily and directly obtained with this optimal trajectory.

In the literature, most of scene text detection approaches are only based on one channel (gray) and one scale (original size), which always result in missing some important characters, especially for text with skew and perspective distortions in scene videos. Here, to deal with the problem of missing characters in scene images (individual video frames), we construct a robust and precise multi-orientation text detection system in scene images by learning locally which can extensively locate possible characters with multi-channel and multi-scale information fusion [Yin *et al.*, 2015; Pei *et al.*, 2016]. In our method, an adaptive multi-channel character grouping method is first proposed to robust extract all possible character candidates, and an Adaboost classifier is then to properly identify character candidates as characters or non-characters. A single-link clustering with distance metric learning is thereafter used to adaptively group characters into text regions, and an effective hybrid filter with Convolution Neural Networks, AdaBoost and Bayesian classifiers is finally designed to precisely verify the extracted text regions.

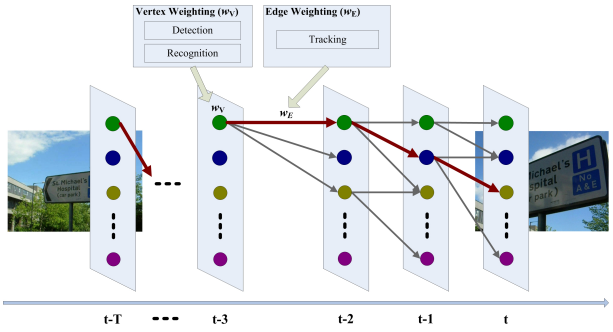

Figure 1: A unified integration framework with dynamic programming for scene text detection in video by learning locally and globally, where w_V and w_E are vertices and edges' weights respectively of the tracking network (a weighted graph), and the RED line is an assumed optimal path by searching globally.

3.2 Tracking Based Text Detection by Learning Globally with Dynamic Programming

In this section, we present a multi-strategy tracking based scene text detection approach. The robust multi-orientation scene text detection method as described above is first used to detect text frame by frame. Then text trajectories are created dynamically when a detected text region in the previous frame does not match any existed text regions in the current frame. For a new trajectory, a tracking by detection method is utilized to verify whether it is valid or not. The trajectory is valid if it is matched in first four consecutive frames; otherwise, it will be regarded as noise and discarded. Afterward, the multi-strategy tracking algorithm which uses tracking by detection, Spatio-Temporal Context Learning (STCL) [Zhang *et al.*, 2013] and template matching is applied to predict the candidate text position respectively in consecutive frames. Finally, a dynamic programming method is used to search globally and decide which candidate text region is the best matching for the target text region. Upon the processing of frame n, the trajectory is removed if there is no matching text region by DCT-based hash algorithm [Zauner, 2010] and the similarity of color histogram.

Text Tracking

Text tracking is to determine the time that text regions appear and disappear, and the location of text continuously in dynamic video frames. Its main aim is to reduce false alarms and improve the accuracy of detection in the dynamic scenes. Three tracking algorithms are used in our multi-strategy tracking method, namely, tracking by detection, STCL and template matching. In the process of creating text trajectory, the tracking-by-detection method associate detected results in successive frames to initialize new trajectories. To reduce false alarms, we assume that the text is regarded as true text if it is detected in four consecutive frames. In the process of tracking the valid text trajectory, the tracking-by-detection method is used to match the tracked outputs in the previous frame and detected outputs in the current frame. The core step in tracking-by-detection is the Hungarian algorithm which maps text regions in successive frames. We assume that the difference of the same text in position and color histogram information are normal distributions with zero means. The product of their probabilities is the input of Hungarian algorithm. The STCL method is used to predict the position in the current frame of the detected or tracked text region in the previous frame in the valid trajectory, which uses Bayesian framework based to model the spatio-temporal relationships between the object of interest and its surrounding regions. Thus, the tracking is to compute a confidence map and obtain the best target location by maximizing an object location likelihood function [Zuo *et al.*, 2015]. Moreover, instead of linear prediction which is not fit for arbitrary motion, a normalized correlation coefficient based template matching method is added to predict the position in the current frame of the detected or tracked text region in the previous frame in the valid trajectory in this paper. The normalized correlation coefficient is,

$$R(x,y) = \frac{\sum_{x',y'} (T'(x',y') \bullet I'(x+x',y+y'))}{\sqrt{\sum_{x',y'} T'(x',y')^2 \bullet \sum_{x',y'} I'(x+x',y+y')^2}} \quad (1)$$

where $R(x,y)$ is the estimated value of a point (x,y) in result image R, I' and T' are image and sliding template respectively, and x' and y' are the position of the sliding template T' in image I'. At last, the location of maximum value in R is the centroid of prediction location.

Since the normalized correlation coefficient based template matching always returns a position even when the matched block does not contain text, the combination of the similarity of color histogram and DCT-based hash algorithm [Zauner, 2010] is used to determine whether the target text region and the predictive text region are similar or not. The trajectory will be eliminated when there is no matching in tracking process.

Here, tracking-by-detection makes use of detection outputs to initialize new trajectories and amend tracking outputs, and usually has high accuracy. Template matching is used to tackle text blur challenges, but failed to handle multi-scale challenges, while STCL is applied to solve the problem of multi-scale text. Both template matching and STCL can generally improve the recall performance. So, our multi-strategy tracking approach and the dynamic programming algorithm can combine advantages of these three tracking techniques and obtain good accuracy with fair recall.

Searching Globally with Dynamic Programming

Instead of the rule-based technique in our previous method [Zuo *et al.*, 2015], we have developed a dynamic programming based method (an iterative improvement over [Zuo *et al.*, 2015]) to globally search and select the best matching from the candidate text regions for the target text. Moreover, in the process of multi-strategy tracking, each detected and tracked text region in the valid trajectory will be regarded as a node added into the network, and all text regions in a frame is corresponded with a layer of the dynamic network (see Figure 1).

Node weights and edge weights are the important part for building a dynamic programming network. In our work, the

combination of the detection confidence score and the recognition respond score [1] are used to compute the weight of the node,

$$W_V(n) = \alpha N_d(n) + \beta N_r(n) \quad (2)$$

where $W_V(n)$ is the sum score of the node n, $N_d(n)$ and $N_r(n)$ are the detection confidence score and the recognition respond score of this node n respectively, α and β are weight coefficients between 0 and 1. The similarity of color histogram combined with edit distance is utilized to compute edge weight between the two nodes in consecutive frames with

$$W_E(n_1, n_2) = \lambda P_{col}(n_1, n_2) + \mu P_{edist}(n_1, n_2) \quad (3)$$

where n_1 and n_2 are two nodes in the consecutive frame, $W_E(n_1, n_2)$ is the sum similarity between the two nodes, $P_{col}(n_1, n_2)$ and $P_{edist}(n_1, n_2)$ are the similarity of color histogram and edit distance of the two nodes respectively, λ and μ are weight coefficient between 0 and 1. The score of each node is the maximum score from the start node to the current node, i.e.,

$$\max Score(n_i^j) = \begin{cases} \max Score(n_{i-1}^k) + W_V(n_i^j) \\ \quad + W_E(n_{i-1}^k, n_i^j), & i > 1 \\ n_1^1, & i = 1 \end{cases} \quad (4)$$

where $maxScore(n_i^j)$ is the score of the jth node in the ith layer, the range of k and j is from 1 to $\max(layer_i, 6)$, and $layer_i$ is the total number of nodes in ith layer.

In the process of the dynamic programming construction, each text region including a detected or tracked text region (predicted text region) in the previous frame will be predicted in the current frame. If the total number of predicted text regions for the same text region are more than six in a frame, we will select the text region with a higher score and the overlap less than 95%. The text region with the less similarity of the color histogram will be discarded when the overlap between the text regions is more than 95%. Finally, a node will be created with $W_V = 0$ when the trajectories are no matched in tracking process. Thus, the nodes with the maximum score are determined as the text positions in their frames.

4 Experiments

4.1 Experiments with Scene Images

For the task of text detection in scene images (scene frames), we choose two public multi-orientation scene text datasets, MSRA-TD500 [Yao *et al.*, 2012] and USTB-SV1K [Yin *et al.*, 2015] , and use the evaluation protocol in Yao's et al's work [Yao *et al.*, 2012].

The MSRA-TD500 dataset is a multi-orientation dataset with 500 images where 300 images is for training and the rest is for testing. We compare our multi-channel fusion method with five state-of-the-art methods: two of Yin et al.'s method [Yin *et al.*, 2014; 2015], two of Yao et al.'s methods [Yao *et al.*, 2012; 2014] and Kang et al.'s method [Kang *et al.*, 2014]. As is shown in Table 2, the method with multi-information fusion strategy performs much better than all other methods. USTB-SV1K is a dataset crawled from Google Street View directly. The images have the lower resolution and are blurred artificially to some degree. So it is a more challenging dataset for text detection. Here, we compared our system with some other advanced methods. The experimental results (see Table 3) are same as the results on MSRA-TD500, namely, the fusion strategy contributes to improve recall and the CNN filter improves precision effectively. However, a part of images are severely distorted in USTB-SV1K. Our text detection method in individual frames is sensitive to perspectively distorted text.

4.2 Experiments with Scene Videos

To evaluate our tracking based detection approach, a public video dataset with a variety of scene videos is first used in our experiments [2]. In these scene videos, some text regions have affected by nature noise, distortion, blurring, hard illumination changes and occlusion. We compare our approach with the methods in [Minetto *et al.*, 2011] and [Zuo *et al.*, 2015]. Here, we use the well-known metrics precision, recall and f-score defined in [Lucas, 2005].

Table 1 presents tracking based scene text detection performance of Minetto et al.'s method [Minetto *et al.*, 2011], the method in [Zuo *et al.*, 2015] and the proposed method in this paper respectively. As seen from the table, with the performance of text detection in [Minetto *et al.*, 2011], the average performance of text detection in [Zuo *et al.*, 2015] has an increase of f-score by 12%. What's more, the performance of tracking based text detection method proposed in this paper is 6% higher than the method in [Zuo *et al.*, 2015]. In other words, the proposed multi-strategy tracking method is effective and robust to reduce false alarms and improve the accuracy of text detection. Figure 2 shows some scene text tracking results by our method on the public dataset.

Moreover, we also perform experiments of our method on the recent challenging dataset of ICDAR 2015 Robust Reading Competition Challenge 3 (Text Detection and Recognition in Scene Videos) [3]. This dataset includes a training set of 25 videos (13450 frames in total) and a test set of 24 videos (14374 frames in total), where are collected by the organizers in different countries, including text in different languages. The video sequences correspond to 7 high level tasks in indoors and outdoors scenarios. Moreover, 4 different cameras are used for capturing different sequences. We use the ICDAR'15 Robust Reading Competition Platform to evaluate our proposed approach, and comparative results are shown in Table 4 [4], where "ATA" is the official metric of the ICDAR 2015 Robust Reading Competition Challenge [Karatzas *et al.*, 2015b]. Our proposed approach has the best performance for video text detection on this competition dataset.

[1] In our system, a CNN-based word recognition technique is used [Jaderberg *et al.*, 2016]. The recognition respond scores are directly derived from the outputs of the classifier.

[2] http://www.liv.ic.unicamp.br/~minetto/datasets/text/VIDEOS/

[3] http://rrc.cvc.uab.es/?ch=3&com=introduction

[4] In Table 4, we present competition results from the top 3 participation teams (each team with only its highest performance). Actually, we won this competition and there were all seven submissions of four teams in the competition.

Table 1: Comparative results for text detection in scene videos (%).

Video	Minetto et al.'s [Minetto *et al.*, 2011]			Zuo et al.'s [Zuo *et al.*, 2015]			Proposed method		
	Precision	Recall	f	Precision	Recall	f	Precision	Recall	f
V1	0.55	0.80	0.63	0.82	0.62	0.71	0.82	0.70	0.76
V2	0.57	0.74	0.64	0.90	0.80	0.85	0.92	0.81	0.86
V3	0.60	0.53	0.56	0.75	0.60	0.67	0.73	0.64	0.68
V4	0.73	0.70	0.71	0.83	0.77	0.80	0.88	0.82	0.85
V5	0.60	0.70	0.63	0.88	0.62	0.72	0.89	0.87	0.88
average	0.61	0.69	0.63	0.84	0.68	0.75	0.85	0.77	**0.81**

Figure 2: Text tracking and detection results for our approach on sample videos.

Table 2: Experimental results in on MSRA-TD500.

Method	Recall	Precision	f-measure
Proposed method	0.5806	**0.9497**	**0.7207**
[Yin *et al.*, 2014]	**0.63**	0.81	0.71
[Yin *et al.*, 2015]	0.61	0.71	0.66
[Yao *et al.*, 2012]	**0.63**	0.63	0.60
[Yao *et al.*, 2014]	0.62	0.64	0.61
[Kang *et al.*, 2014]	0.62	0.71	0.66

Table 3: Experimental results on USTB-SV1K.

Method	Recall	Precision	f-measure
Proposed method	0.488	**0.5369**	**0.5112**
[Yin *et al.*, 2014]	0.4541	0.4985	0.4753
[Yin *et al.*, 2015]	0.4518	0.45	0.4509
[Yao *et al.*, 2014]	0.4405	0.458	0.4491

ATA (Average Tracking Accuracy) provides an object tracking measure that penalizes fragments both in temporal and spatial dimensions. MOTA (Multiple Object Tracking Accuracy) is another metric to evaluate the performance of object tracking, which penalizes more on false positive and mismatch. MOTP (Multiple Object Tracking Precision) is used to evaluate the detection performance without the explicit penalization on the temporal errors. The highest score of ATA achieved by our method means that our method retrieves more tracks than others. The lower scores of MOTA show that because of severely perspective and aligned distortions, the number of false positives of our method is relatively large, which is also a near future topic of our research.

The main limitation of the proposed method is the tracking degradation in presence of severe motion blur. In addition, a fast and effective text detection method is important in such a framework to initialize the text target and ensure that tracking does not deteriorate rapidly. As future work, on the one hand, the text detection algorithm will be optimized further to reduce the delay on initial text target and provide robustness to the tracking module. On the other hand, the tracking algorithms will be extended by other tracking techniques to improve robustness for some challenges such as severe motion blur and occlusions.

Table 4: Experimental results (%) on the dataset of ICDAR 2015 Robust Reading Competition Challenge 3 (Text Detection in Scene Videos), where "AJOU" is the 2^{nd} winning group, "StradVision" is the 3^{rd} winning group and, "Deep2Text I" is the method name of our partition team in the competition.

Method	MOTP	MOTA	ATA
Proposed method (Deep2Text I)	71.01	40.77	**45.18**
AJOU	73.25	53.45	38.77
StradVision	70.82	47.58	32.12

5 Conclusion

In this paper, we construct a robust and precise text detection system in scene videos by locating character candidates extensively (learning locally) and searching text regions globally (learning globally). With the multi-orientation scene text detection method in images (video frames), we proposed a multi-strategy tracking based text detection approach in scene videos to globally search and select the best text region with dynamic programming. Experiments on a variety of public datasets verify our proposed methods. Impressively, our proposed technology won the ICDAR 2015 Robust Reading Competition (video text detection).

Acknowledgments

The research is partly supported by National Natural Science Foundation of China (61473036).

References

[Aoki *et al.*, 1999] Hisashi Aoki, Bernt Schiele, and Alex Pentland. Realtime personal positioning system for a wearable computer. In *Proceedings of The Third International Symposium on Wearable Computers*, pages 37–43, 1999.

[Chen and Yuille, 2004] Xiangrong Chen and A.L. Yuille. Detecting and reading text in natural scenes. In *Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR'04)*, pages 366–373, 2004.

[Ezaki *et al.*, 2004] Nobuo Ezaki, Marius Bulacu, and Lambert Schomaker. Text detection from natural scene images: Towards a system for visually impaired persons. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)*, volume 2, pages 683–686, 2004.

[Fragoso *et al.*, 2011] Victor Fragoso, Steffen Gauglitz, Shane Zamora, Jim Kleban, and Matthew Turk. Translatar: A mobile augmented reality translator. In *Proceedings of the 2011 IEEE Workshop on Applications of Computer Vision (WACV'11)*, pages 497–502, 2011.

[Gomez and Karatzas, 2014] Llifs Gomez and Dimosthenis Karatzas. Mser-based real-time text detection and tracking. In *Proceedings of the 22nd International Conference on Pattern Recognition (ICPR'14)*, pages 3110–3115, 2014.

[Goto and Tanaka, 2009] Hideaki Goto and Makoto Tanaka. Text-tracking wearable camera system for the blind. In *Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR'09)*, pages 141–145, 2009.

[Haritaoglu, 2001] Ismail Haritaoglu. Scene text extraction and translation for handheld devices. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'01)*, volume 2, pages 408–413, 2001.

[Iwatsuka *et al.*, 2004] Kentaro Iwatsuka, Kazuhiko Yamamoto, and Kunihito Kato. Development of a guide dog system for the blind people with character recognition ability. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)*, volume 1, pages 453–456, 2004.

[Jaderberg *et al.*, 2016] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision*, 116(1):1–20, 2016.

[Kang *et al.*, 2014] Le Kang, Yi Li, and David Doermann. Orientation robust textline detection in natural images. In *Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR'14)*, 2014.

[Karatzas *et al.*, 2015a] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, and etc. ICDAR 2015 Competition on Robust Reading. In *Proceedings of ICDAR*, pages 1156–1160, 2015.

[Karatzas *et al.*, 2015b] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, and etc. ICDAR 2015 Competition on Robust Reading. In *Proceedings of ICDAR*, pages 1156–1160, 2015.

[Li and Doermann, 1998] Huiping Li and David Doermann. Automatic text tracking in digital videos. In *Proceedings of the 1998 IEEE Second Workshop on Multimedia Signal Processing*, pages 21–26, 1998.

[Li *et al.*, 2000] Huiping Li, David Doermann, and Omid Kia. Automatic text detection and tracking in digital video. *IEEE Trans. Image Processing*, 9(1):147–156, 2000.

[Liu and Wang, 2012] Xiaoqian Liu and Weiqiang Wang. Robustly extracting captions in videos based on stroke-like edges and spatio-temporal analysis. *IEEE Trans. Multimedia*, 14(2):482–489, 2012.

[Lucas, 2005] Simon M Lucas. ICDAR 2005 text locating competition results. In *Proceedings of ICDAR*, pages 80–84, 2005.

[Merino-Gracia and Mirmehdi, 2007] Carlos Merino-Gracia and Majid Mirmehdi. A framework towards realtime detection and tracking of text. In *Proceedings of the Second International Workshop on Camera-Based Document Analysis and Recognition (CBDAR'07)*, pages 10–17, 2007.

[Merino-Gracia and Mirmehdi, 2014] Carlos Merino-Gracia and Majid Mirmehdi. Real-time text tracking in natural scenes. *IET Computer Vision*, 8(6):670–681, 2014.

[Merino-Gracia *et al.*, 2012] Carlos Merino-Gracia, Karel Lenc, and Majid Mirmehdi. A head-mounted device for recognizing text in natural scenes. In *Camera-Based Document Analysis and Recognition*, pages 29–41. 2012.

[Mi *et al.*, 2005] Congjie Mi, Yuan Xu, Hong Lu, and Xiangyang Xue. A novel video text extraction approach based on multiple frames. In *Proceedings of the Fifth International Conference on Information, Communications and Signal*, pages 678–682, 2005.

[Minetto *et al.*, 2011] Rodrigo Minetto, Nicolas Thome, Matthieu Cord, Neucimar J Leite, and Jorge Stolfi. Snooopertrack: Text detection and tracking for outdoor videos. In *Proceedings of the 18th IEEE International Conference on Image Processing (ICIP'11)*, pages 505–508, 2011.

[Na and Wen, 2010] Yinan Na and Di Wen. An effective video text tracking algorithm based on sift feature and geometric constraint. In *Advances in Multimedia Information Processing-PCM 2010*, pages 392–403. 2010.

[Nguyen *et al.*, 2014] Phuc Xuan Nguyen, Kai Wang, Serge Belongie, and Cornell NYC Tech. Video text detection and recognition: Dataset and benchmark. *Applications of Computer Vision (WACV)*, 2014.

[Pei *et al.*, 2016] Wei-Yi Pei, Chu Yang, and Xu-Cheng Yin. Multi-orientation scene text detection with multiple information fusion. In *International Conference on Pattern Recognition (ICPR'16)*, 2016. submitted.

[Petter *et al.*, 2011] Marc Petter, Victor Fragoso, Matthew Turk, and Charles Baur. Automatic text detection for mobile augmented reality translation. In *Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV'11)*, pages 48–55, 2011.

[Sanketi *et al.*, 2011] Pannag Sanketi, Huiying Shen, and James M Coughlan. Localizing blurry and low-resolution text in natural images. In *Proceedings of the 2011 IEEE Workshop on Applications of Computer Vision (WACV'11)*, pages 503–510, 2011.

[Shi and Xu, 2005] Xi Shi and Yangsheng Xu. A wearable translation robot. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA'05)*, pages 4400–4405, 2005.

[Shiratori *et al.*, 2006] Hiroki Shiratori, Hideaki Goto, and Hiroaki Kobayashi. An efficient text capture method for moving robots using dct feature and text tracking. In *Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)*, volume 2, pages 1050–1053, 2006.

[Shivakumara *et al.*, 2012] Palaiahnakote Shivakumara, Rushi Padhuman Sreedhar, Trung Quy Phan, Shijian Lu, and Chew Lim Tan. Multioriented video scene text detection through bayesian classification and boundary growing. *IEEE Trans. Circuits and Systems for Video Technology*, 22(8):1227–1235, 2012.

[Shivakumara *et al.*, 2013] Palaiahnakote Shivakumara, Trung Quy Phan, Shijian Lu, and Chew Lim Tan. Gradient vector flow and grouping-based method for arbitrarily oriented scene text detection in video images. *IEEE Trans. Circuits and Systems for Video Technology*, 23(10):1729–1739, 2013.

[Tanaka and Goto, 2007] Makoto Tanaka and Hideaki Goto. Autonomous text capturing robot using improved dct feature and text tracking. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR'07)*, volume 2, pages 1178–1182, 2007.

[Tanaka and Goto, 2008] Makoto Tanaka and Hideaki Goto. Text-tracking wearable camera system for visually-impaired people. In *Proceedings of the 19th International Conference on Pattern Recognition (ICPR'08)*, pages 1–4, 2008.

[Wan and Van Der Merwe, 2000] Eric A Wan and Rudolph Van Der Merwe. The unscented kalman filter for nonlinear estimation. In *Proceedings of The IEEE Adaptive Systems for Signal Processing, Communications, and Control Symposium (AS-SPCC'00)*, pages 153–158, 2000.

[Wang and Wei, 2010] Zhen Wang and Zhiqiang Wei. An efficient video text recognition system. In *Proceedings of the 2nd International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC'10)*, volume 1, pages 174–177, 2010.

[Wang *et al.*, 2013] Baokang Wang, Changsong Liu, and Xiaoqing Ding. A research on video text tracking and recognition. In *Proceedings of IS&T/SPIE Electronic Imaging*, pages 86640G–86640G, 2013.

[Wu *et al.*, 2004] Wen Wu, Xilin Chen, and Jie Yang. Incremental detection of text on road signs from video with application to a driving assistant system. In *Proceedings of the 12th annual ACM International Conference on Multimedia (ACM MM'04)*, pages 852–859, 2004.

[Wu *et al.*, 2005] Wen Wu, Xilin Chen, and Jie Yang. Detection of text on road signs from video. *IEEE Trans. Intelligent Transportation Systems*, 6(4):378–390, 2005.

[Yao *et al.*, 2012] Cong Yao, Xiang Bai, Wenyu Liu, Yi Ma, and Zhuowen Tu. Detecting texts of arbitrary orientations in natural images. In *Proceeding of CVPR*, pages 1083–1090, 2012.

[Yao *et al.*, 2014] C. Yao, X. Bai, and W. Liu. A unified framework for multi-oriented text detection and recognition. *IEEE Trans. Image Processing*, 23(11):4737–4749, 2014.

[Ye and Doermann, 2015] Qixiang Ye and David Doermann. Text detection and recognition in imagery: A survey. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 37(7):1480–1500, 2015.

[Yin *et al.*, 2014] Xu-Cheng Yin, Xuwang Yin, Kaizhu Huang, and Hong-Wei Hao. Robust text detection in natural scene images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 36(5):970–983, 2014.

[Yin *et al.*, 2015] Xu-Cheng Yin, Wei-Yi Pei, Jun Zhang, and Hong-Wei Hao. Multi-orientation scene text detection with adpative clustering. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 37(9):1930–1937, 2015.

[Yin *et al.*, 2016] Xu-Cheng Yin, Ze-Yu Zou, Shu Tian, and Cheng-Lin Liu. Text detection, tracking and recognition in video: A comprehensive survey. *IEEE Trans. Image Processing*, 2016. accepted.

[Yusufu *et al.*, 2013] Tuoerhongjiang Yusufu, Yiqing Wang, and Xiangzhong Fang. A video text detection and tracking system. In *Proceedings of the 2013 IEEE International Symposium on Multimedia (ISM'13)*, pages 522–529, 2013.

[Zauner, 2010] Christoph Zauner. *Implementation and benchmarking of perceptual image hash functions*. 2010.

[Zhang *et al.*, 2013] H. Zhang, K. Zhao, Y.-Z. Song, and J. Guo. Text extraction from natural scene images: A survey. *Neurocomputing*, 122:310–323, 2013.

[Zuo *et al.*, 2015] Ze-Yu Zuo, Shu Tian, and Xu-Cheng Yin. Multi-strategy tracking based text detection in scene videos. In *Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR'15)*, 2015.