Guided Text Spotting for Assistive Blind Navigation in Unfamiliar Indoor Environments

Xuejian Rong[†] Bing Li[†] J. Pablo Muñoz[§]

Jizhong Xiao[†§] Aries Arditi[ℓ] Yingli Tian[†§]

[†]The City College, City University of New York, NY
{xrong,bli,jxiao,ytian}@ccny.cuny.edu
[§]The Graduate Centor, City University of New York, NY
jmunoz2@gradcenter.cuny.edu
[ℓ]Visibility Metrics LLC, Chappaqua, NY
arditi@visibilitymetrics.com

Abstract. Scene text in indoor environments usually preserves and communicates important contextual information which can significantly enhance the independent travel of blind and visually impaired people. In this paper, we present an assistive text spotting navigation system based on an RGB-D mobile device for blind or severely visually impaired people. Specifically, a novel spatial-temporal text localization algorithm is proposed to localize and prune text regions, by integrating stroke-specific features with a subsequent text tracking process. The density of extracted text-specific feature points serves as an efficient text indicator to guide the user closer to text-likely regions for better recognition performance. Next, detected text regions are binarized and recognized by off-the-shelf optical character recognition methods. Significant non-text indicator signage can also be matched to provide additional environment information. Both recognized results are then transferred to speech feedback for user interaction. Our proposed video text localization approach is quantitatively evaluated on the ICDAR 2013 dataset, and the experimental results demonstrate the effectiveness of our proposed method.

1 Introduction

Texts in natural scenes matter, since they usually convey significant semantic information and often serve as effective cues in unfamiliar environments for wayfinding. According to the World Health Organization[1], there are more than 39 million legally blind and 285 million visually impaired people living across the world, and this number is still growing at an alarming rate. Although many personal Text-to-Speech assistive systems [1] have been developed for recognizing product labels, grocery signs, indoor indicators, and currency and bills, effective scene text spotting (including text detection and recognition) from videos captured by mobile devices in natural scenes remains a challenging problem.

[1] http://tinyurl.com/who-blindness

Fig. 1. Samples which demonstrate the small size and relatively low resolution of many interesting text regions with respect to the large scale of the whole scene image.

In recent years, the data collected from mobile smartphones and wearable devices has become increasingly important for a broad range of applications, including static Photo Optical Character Recognition (OCR) and dynamic Video OCR. To extract text information in complex natural scenes, effective and efficient scene text detection and recognition algorithms are essential. However, extracting scene text from mobile devices is challenging due to 1) cluttered backgrounds with noise, blur, and non-text background outliers, such as grids and bricks; 2) diversity of text patterns such as script types, illumination variation, and font size; and 3) the limitations of mobile devices such as limited computational capability, lower image/video resolution, and restricted memory.

In spite of these challenges, many text spotting (from text localization to word recognition) approaches have been recently developed and demonstrated effectiveness in different applications [2,3,4,5,6]. In practice, *Google Translate* and *Microsoft Translator* applications on iOS and Android platforms have been widely used to translate text in photos to a readable sentence in other languages to help foreign tourists, but similar applications based on videos on mobile devices still remain to be explored. On the one hand, simply applying current photo-based text spotting methods to individual frames ignores the continuous temporal cues in consecutive frames. On the other hand, the photo-based text detection and recognition process is usually time-consuming and doesn't meet the efficiency requirement of mobile devices. Moreover, the recognition process of detected text regions often consumes the most computation time in the end-to-end text spotting process [4], and inevitably suffers from the tiny text regions extracted from the large natural scene image.

Considering all the above limitations, we here propose a guided text spotting approach that reduces the number of text recognition steps in continuous

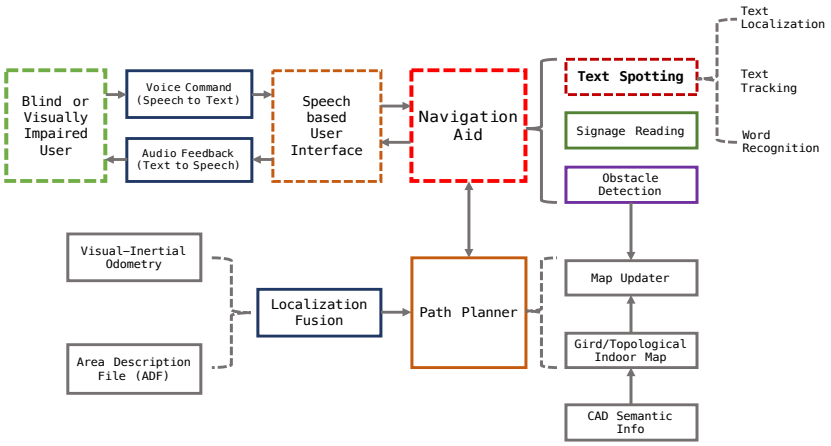

Fig. 2. Flowchart of the proposed Intelligent Situation Awareness and Navigation Aid system including the text spotting modules.

videos frames, and gradually guides the user to move closer to the preliminarily detected text regions for better recognition performance. Specifically, in the initial text localization step, a stroke-specific feature detector tuned for lower resolution videos and computation requirement is implemented to quickly propose candidate text regions in natural scene frames. The candidate text regions are then tracked based on the feature points across consecutive video frames to reduce average computational load, eliminate occasional false alarms, and guide the blind user to aim the camera on the mobile device to the most likely text regions. If a text region has been localized, verified, and tracked for a sufficient number of subsequent frames, it is considered as successfully detected as the primary text region. Afterward, an off-the-shelf text recognition approach [7] is applied to translate the scene text into meaningful word strings. The text detection and final text recognition results are passed to the text-to-speech engine to generate voice guidance information for blind users.

Due to the importance and usefulness of many signage indicators (text and non-text (see Fig. 1) existing in the blind navigation environments, we also present a template-matching based approach for extracting the signs to provide more semantic information besides the text spotting process. To demonstrate the effectiveness of the proposed methods in a real blind navigation application, an obstacle-aware **assistive wearable indoor navigation system** is designed and presented, including a speech-based user interaction interface.

The rest of the paper is organized as follows: in Sec. 2, an overview of existing assistive navigation and text spotting methods is presented. Sec. 3 describes the main components of the proposed indoor navigation system. Sec. 4 introduces the proposed signage reading method, the video text localization and tracking approach, and the speech-based user interaction interface. Sec. 5 presents the experimental results. Sec. 6 describes our conclusions.

2 Related Work

Wearable Indoor Navigation Systems. In recent years, there have been numerous efforts to develop electronic travel aids (ETA) [8] to improve the orientation and mobility of the visually impaired. Most ETAs are designed to improve and enhance independent travel, rather than to replace conventional aids such as the guide dog or long cane.

Various ETAs including different kinds of sensors have been proposed [9,10,11], which usually have in common three basic components: a sensor, a processing system, and an interface. A sensor captures data from the environment in a specific type. The data are then processed to generate useful information for the visually impaired user. Lastly, an interface delivers the processed information to the user using an appropriate sensory modality such as auditory or tactile to convey information. We refer the reader to [9] for a more complete review of the recent development of wearable indoor navigation systems.

Text Spotting in the Wild. Although most of the existing scene text spotting approaches focus on text detection and recognition from a single high-resolution image, some methods have been proposed for text detection in video [3,12,13]. These methods can be briefly divided into connected component-based, texture-based, edge and gradient-based methods [14]. Since connected component-based methods [6] require character shapes, they may not achieve good accuracies for low-resolution text images with complex backgrounds. To handle complex backgrounds, texture feature-based methods have been developed [15]. These methods are computationally expensive and their performance depends on the number of trained classifier and collected samples. To improve efficiency, edge and gradient-based methods have been proposed [2]. These methods are efficient but more sensitive to cluttered backgrounds and hence produce more false positives. However, most of these methods are not able to suit the mobile computational capability and still rely on individual frames instead of utilizing temporal information of video stream.

3 Indoor Navigation System

Before introducing our proposed guided text spotting methods in detail, we first give an overview of the Intelligent Situation Awareness and Navigation Aid system in which we implemented them. The hardware (shown in Fig. 3) comprises of a chest-mounted mobile device (Google Tango Tablet[2]) with an integrated RGB-D camera. The software consists of our algorithms for navigation, scene text spotting, scene signage reading, and speech based user interface which are all developed and deployed on the Google Tango device. The main components of the software architecture are shown in Fig. 2.

[2] `https://get.google.com/tango`

Fig. 3. The basic hardware configuration of the proposed assistive navigation system, including the Google Tango Tablet device and the 3D printed chest level tablet mount.

Initialization and Mapping. We use the Google Tango Android tablet device for our prototype design mainly due to its portability, its ability to build 3D sparse feature maps called Area Description File (ADF), and its ability to localize based on the ADF. A feature-based Simultaneous Localization and Mapping (SLAM) module running on the Tango device provides a feature model as an ADF map for area learning and area recognition. First, the model file is parsed and geometric objects such as texts as room labels, ellipses as doors, polylines as contours are acquired. Then semantic topological graph connections between room labels and doors are analyzed using region growing algorithm, and semantic landmarks and areas of interest are updated into the semantic map. Finally, a multi-floor semantic map is built as the graph between common connectors such as stairs, elevator, and escalator.

Navigation with Obstacle Avoidance. Based on the indoor semantic map, we create a node-edge based graph and then use it to perform indoor assistive navigation for the blind user. A graph searching algorithm is applied to generate an optimal path from the current position node to the node nearby the specified destination. Afterward, a waypoint representation of the route is refined from the path and delivered to the user for guidance. The proposed system further provides the user local obstacle direction information such as front/front right/head-height obstacles. Finally, in the scenarios where there are multiple obstacles, obstacle position, and size information are updated into the global map, and a request is set for path planning to generate a new path. The obstacle projection in 3D space is illustrated in Fig. 4.

Speech Recognition based User Input Interface. During the navigation process, the speech recognition interface, developed on the CMU Sphinx library, keeps working in the background to receive speech commands from the user. The commands include but are not limited to, starting, ending, pausing, resuming, stopping, and restarting the navigation processing. The effectiveness of the

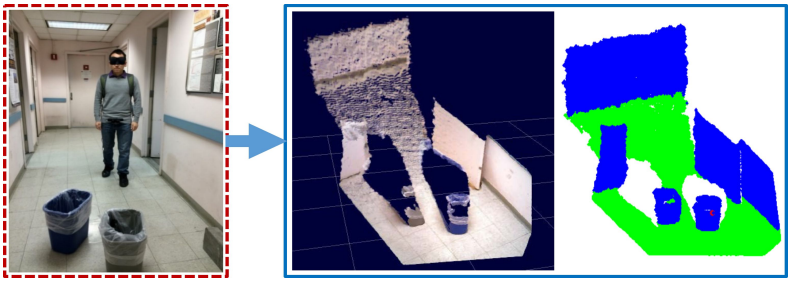

Fig. 4. Demonstration of the 3D projection of the obstacles in front of the user with RGB-D camera, mounted just below his waist. The two images on the right are from the point of view of the camera. The blue pixels represent the corresponding obstacle regions, and the green pixels repesent the ground regions respectively. The red pixels represent the region of the occasionally missing data.

speech to text modules has been validated in practice and proven to effectively boost the practicability of our proposed pipeline in the blind navigation system.

4 Signage and Scene Text Reading for a Navigation Aid

In this section, we focus on describing the signage and scene text spotting approaches in details, including the localization, verification, fusion, and recognition stages. The speech-based interface is also introduced in Sec. 4.3.

4.1 Signage Reading

To effectively recognize the signage most significant for wayfinding and safety in indoor environments, a template matching method is developed to recognize predefined meaningful signage based on the binary online learned descriptor [16]. In practice, we follow a matching-by-detection mechanism. An instance-specific detector is trained based on the pre-collected indicator sign dataset, but it is not updated online to avoid the influence of various training examples, which effectively alleviates the problem of weak binary tests. As in [17], we create a classification system based on a set of N simple binary features of intensity differences, similar to the ones of Binary Robust Independent Elementary Features (BRIEF). Following a sliding window manner, which is common among state-of-the-art detectors, each window candidate is classified as a target sign or as background. The recognized sign is then vocalized via the text-to-speech module (See Sec. 4.3).

4.2 Scene Text Detection

Text Region Localization. Typically, OCR methods present low recognition performance when the texts in the image suffer perspective distortion or are not

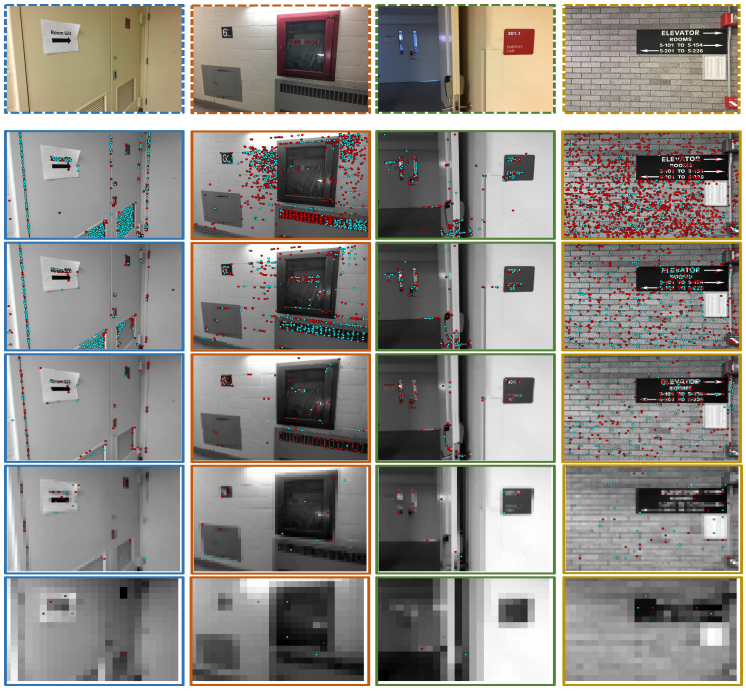

Fig. 5. Each column demonstrates the detected stroke specific features (Red for stroke ending point and Cyan for stroke bend point) on the gradually coarser levels of image pyramid.

properly aligned, centered, scaled or illuminated. This often occurs in ego-centric images or videos captured by a wearable camera in indoor navigation environments. In this case, the OCR performance could be boosted if we could automatically obtain regions of interest containing text and process them to avoid these issues from general scene images. Considering the application needs to run on a mobile device, we start by restricting this initial processing to repetitively detected text regions across consecutive video frames to improve efficiency.

In the proposed video-based text spotting module, a stroke specific text detector, FASText [4], is employed to initially localize the potential text regions, since it is fast, scale and rotation invariant, and usually produces fewer false detections than the detectors commonly used by prevailing scene text localization methods. Considering the observation that almost all the script texts are formed of strokes, stroke keypoints are efficiently extracted and segmented subsequently. General corner detection methods [18] could successfully detect the corners and stroke endings of certain letters such as the letter "K" or "I", but would usually fail on characters whose strokes do not have a corner or an ending such as the letter "O" or the digit "8". In comparison, the FASText detector tends to boost the detection performance of the proposed pipeline by focusing on the detection of stroke ending/bending keypoint at multiple scales. And the keypoints are de-

tected in an image scale-space to allow detection of wider strokes. Each level of the image pyramid is calculated from the previous one by reducing the image size by the scaling factor. A simple non-maximum suppression is also performed on a 3×3 neighborhood to further eliminate the number of the detected feature keypoints (See Fig. 5).

After the initial keypoints have been detected, an efficient Gentle AdaBoost classifier [19] is applied to reduce the still relatively high false detection rate, and eliminate regions which do not correspond to text fragments, including a part of a character, a single character, a group of characters, and a whole word. The classifier exploits features already calculated in the detection phase and an effectively approximated strokeness feature, which plays an important role in the discrimination between text fragments and background clutter. The classification step also accelerates the processing in the subsequent steps. Finally, an efficient text clustering algorithm based on text direction voting is implemented to aggregate detected regions into text line structures and to allow processing by subsequent tracking and recognition. In this step, the unordered set of FASText regions classified as text fragments is clustered into ordered sequences, where each cluster (sequence) shares the same text direction in the image. In other words, individual characters (or groups of characters or their parts) are clustered together to form lines of text.

Although the original FASText detector outperforms many previous text detection approaches on efficiency (average 0.15s on the ICDAR 2013 dataset [20] on a standard PC), it is still not fast enough on the portable platforms without specific tweaking. To make the FASText detector work for mobile computation platforms, we follow the basic structure and feature design in the implementation and tune the detector parameters including the circle size and margin. We also lower the computational load by limiting the maximum number of keypoints per image and reducing the pyramid layers whilst keeping a comparable detection rate for subsequent processing steps.

Verification and Tracking for preliminarily extracted text regions. After the candidate text regions have been proposed by the detector, we further filter the proposals by scene text tracking in order to further reduce the number of candidates which will be processed by the subsequent relatively computation-demanding text recognition. Each frame of the mobile video stream is processed independently and text features from consecutive frames are aggregated. If the same text region in approximately the same location of the image has been tracked across a sufficient number of frames, it is considered as truly detected, and then passed to the following recognition step. The fused density of the preliminarily detected stroke features is also exploited for indicating the most interesting text regions, as illustrated in Fig. 6. The direction guidance information (speech and alert sounds) is generated accordingly to help the blind user to approach the potential text regions to capture the higher resolution images for better text recognition results.

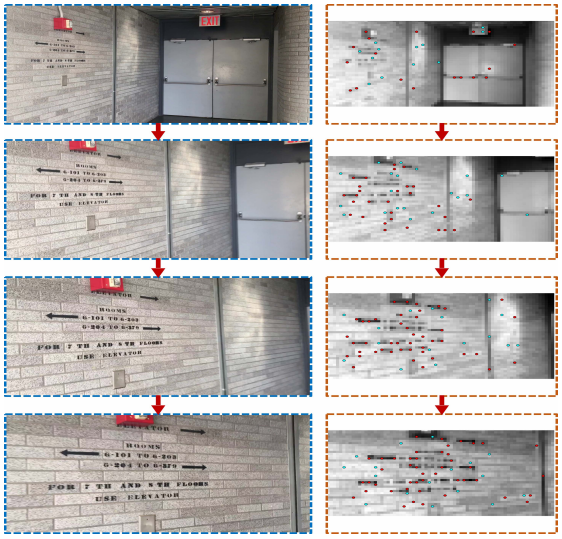

Fig. 6. Demonstration of our text tracking algorithm on consecutive video frames, with the blind user guided by the audio feedback. Density of the previously detected stroke feature points serves as the text-specific indicator and guide the blind or visually impaired user to aim the device to the most likely text regions for better text detection and recognition results.

Unlike previous text tracking algorithms [3,12,21,22], for simultaneously tracking several scene text regions belonging to different words, we apply the multi-object tracking model based on the particle filter in the system implementation, which is capable of handling the variations of lighting and appearance. To avoid challenges of multi-object tracking, three constraints are applied based on our observation. First, the estimation of the scene text character trajectories is not necessary for the same word independently because we can instead estimate the trajectory of the whole text region at first as a hint. Second, the scene text characters within the same word are usually well aligned and are relatively independent of one another. Third, the relative locations of characters are stable. Therefore the inter-object occlusions rarely occur as long as the whole text region is clearly captured. Therefore, we drastically reduce false alarms and boost the efficiency of the whole system.

4.3 Word Recognition and Scene Text to Speech Feedback

Based on the analysis of the most commonly used open source OCR approaches in [23], we decided to use the best compromise option, Tesseract[3], to implement the final mobile navigation prototype. The OCR process could generate better performance if text regions are first extracted and refined by the proposed video

[3] `https://github.com/tesseract-ocr`

text localization algorithm, and then binarized to segment text characters from the background. After completing the sign matching, and the text detection, tracking and recognition process, we further implement the signage and scene text to speech module to convey the results to blind users, including the information of the door numbers, corridor direction, and etc. The built-in speech synthesis engine[4] of Android is adopted in our system to transform the recognized signage and text information to voice output, which provides adaptive navigation support to the blind users.

5 Experimental Results

The proposed system was evaluated on the standard video text spotting benchmark: ICDAR 2013 [20]. The test set of the ICDAR 2013 Robust Reading (Challenge 3) Dataset consists of 15 videos, and the evaluation objective is to localize and track all words in the video sequences. There are many challenging text instances in the dataset (reflections, illumination variations, text written on cluttered backgrounds, textures which resemble characters), but on the other hand, the text is English only and mostly horizontal.

In our experiments, we compared the text tracking results of the proposed method with several state-of-the-art text tracking methods. The evaluation measures consist of Mean Tracking Precision (MOTP), Mean Tracking Accuracy (MOTA), and Average Tracking Accuracy (ATA). More details of these measures are described in [20]. Specifically, Zhao *et al.* and Wu *et al.* adopt the Kanade Lucas Tomasi (KLT) tracker which is not robust to illumination variation across consecutive frames. The performance of [24] heavily relies on the detection results and cannot handle the spatial translation of text regions very well. Mosleh et al. employ Camshift for text region tracking. The implementation details of TextSpotter are proprietary but its performance is reported in [20].

Table 1. Performance of the Proposed and Existing Techniques on Tracking Data of ICDAR 2013 Video Text Dataset.

Method	MOTP	MOTA	ATA
Proposed	0.65	0.39	0.24
Wu et al. [22]	0.61	0.46	0.29
TextSpotter [20]	0.67	0.27	0.12
Mosleh et al. [25]	0.45	0.13	0.03
Li et al. [24]	0.21	0.15	0.07
Zhao et al. [26]	0.24	0.11	0.05

[4] `http://tinyurl.com/android-tts`

As illustrated in Table 1, the effectiveness of the proposed method is comparable with TextSpotter at MOTP, and Wu *et al.* at MOTA and ATA, Since the parameters of the proposed method are tuned to be able to run on a mobile platform with losing accuracy, there is a scope for migrating and comparing all the methods on mobile devices in a real-time environment for more fair evaluation. As to the text detection procedure, the main guidance failure for the text detection process is due to low image contrast, missing threshold in the intensity channel, characters very close to each other, and text-likely textures (see Fig. 7).

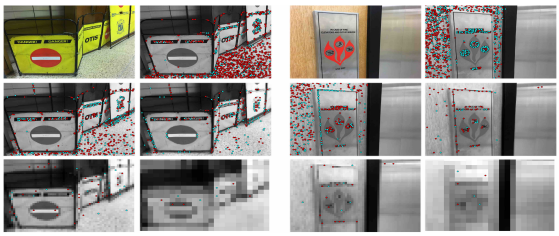

Fig. 7. Guidance difficulty of text localization process caused by low image contrast and text likely textures. Best viewed in color.

6 Conclusion and Future Work

In this paper, we have demonstrated the feasibility of a signage and scene text to speech module as implemented in an assistive wearable indoor navigation system on a Google Tango Tablet device, for better navigation aid to visually impaired users. Our future work will focus on investigating more efficient deep learning based text spotting methods to further boost system performance.

Acknowledgements

This work was supported in part by U.S. Federal Highway Administration (FHWA) grant DTFH 61-12-H-00002, National Science Foundation (NSF) grants CBET-1160046, EFRI-1137172 and IIP-1343402, National Institutes of Health (NIH) grant EY023483.

References

1. Xiong, B., Grauman, K.: Text detection in stores using a repetition prior. WACV (2016)
2. Qin, S., Manduchi, R.: A fast and robust text spotter. WACV (2016)
3. Yin, X., Zuo, Z., Tian, S., Liu, C.: Text detection, tracking and recognition in video: A comprehensive survey. IEEE Trans. on Image Processing (2016)

4. Busta, M., Neumann, L., Matas, J.: Fastext: Efficient unconstrained scene text detector. ICCV (2015)
5. Jaderberg, M., Vedaldi, A., Zisserman, A.: Deep features for text spotting. ECCV (2014)
6. Yin, X., Yin, X., Huang, K., Hao, H.: Robust text detection in natural scene images. IEEE Trans. on Pattern Analysis and Machine Intelligence (2014)
7. Rakshit, S., Basu, S.: Recognition of handwritten roman script using tesseract open source ocr engine. arXiv.org (2010)
8. Muñoz, J.P., Xiao, J.: Detecting and recognizing signage for blind persons to access unfamiliar environments. Network Modeling Analysis in Health Informatics and Bioinformatics (2013)
9. Lees, Y., Medioni, G.: Rgb-d camera based wearable navigation system for the visually impaired. Computer Vision and Image Understanding (2016)
10. Li, B., Muñoz, J.P., Rong, X., Xiao, J., Tian, Y., Arditi, A.: Isana: Wearable context-aware indoor assistive navigation with obstacle avoidance for the blind. ECCV Workshop (2016)
11. Li, B., Zhang, X., Muñoz, J.P., Xiao, J., Rong, X., Tian, Y.: Assisting blind people to avoid obstacles: An wearable obstacle stereo feedback system based on 3d detection. IEEE International Conference on Robotics and Biomimetics (ROBIO) (2015)
12. Rong, X., Yi, C., Yang, X., Tian, Y.: Scene text recognition in multiple frames based on text tracking. IEEE International Conference on Multimedia and Expo (2014)
13. Rong, X., Yi, C., Tian, Y.: Recognizing text-based traffic guide panels with cascaded localization network. ECCV Workshop (2016)
14. Yi, C., Tian, Y.: Text string detection from natural scenes by structure-based partition and grouping. IEEE Transaction on Image Processing (2011)
15. Yi, C., Tian, Y., Arditi, A.: Portable camera-based assistive text and product label reading from hand-held objects for blind persons. IEEE Trans. on Mechatronics (2014)
16. Balntas, V., Tang, L., Mikolajczyk, K.: Bold - binary online learned descriptor for efficient image matching. CVPR (2015)
17. Ozuysal, M., Calonder, M., Lepetit, V., Fua, P.: Fast keypoint recognition using random ferns. IEEE Trans. on Pattern Analysis and Machine Intelligence (2010)
18. Rosten, E., Drummond, T.: Fusing points and lines for high performance tracking. ICCV (2005)
19. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting. The Annals of Statistics (2000)
20. Karatzas, D.: Icdar 2013 robust reading competition. ICDAR (2013)
21. Goto, H., Tanaka, M.: Text-tracking wearable camera system for the blind. ICDAR (2009)
22. Wu, L., Shivakumara, P., Lu, T.: A new technique for multi-oriented scene text line detection and tracking in video. IEEE Trans. on Multimedia (2015)
23. Cambra, A., Murillo, A.: Towards robust and efficient text sign reading from a mobile phone. (2011)
24. Li, H., Doermann, D., Kia, O.: Automatic text detection and tracking in digital video. IEEE Trans. on Image Processing (2000)
25. Mosleh, A., Bouguila, N., Hamza, A.: Automatic inpainting scheme for video text detection and removal. IEEE Trans. on Image Processing (2013)
26. Zhao, X., Lin, K., Fu, Y., Hu, Y., Liu, Y.: Text from corners: a novel approach to detect text and caption in videos. IEEE Trans. on Image Processing (2011)