Person Re-identification with End-to-End Scene Text Recognition

Kamlesh, Pei Xu(✉), Yang Yang, and Yongchao Xu

The School of Electronic Information and Communications,
Huazhong University of Science and Technology (HUST), Wuhan 430074, China
kdnarwani@hotmail.com, xupei_chn@163.com, {yangzw,yongchaoxu}@hust.edu.cn

Abstract. Person re-identification (Re-ID) has become increasingly popular in vision community. Many previous works rely on either feature representation learning and/or metric learning. Different from classical methods, we find that some text in images could be considered as a key cue for differentiating persons under some circumstances (e.g., racing bib number of marathon participants). Based on this observation, we propose to simplify the person Re-ID problem into an end-to-end text recognition problem. Thanks to many powerful state-of-the-art text recognition systems, we can largely improve the efficiency and accuracy of person re-identification in such circumstances. Moreover, we collect a dataset consisting of 9706 marathon images and propose an appropriate measurement to benchmark person identification. Our work provides a promising perspective to person Re-ID and end-to-end text recognition fields, showing also high potentials for video surveillance and image retrieval.

Keywords: Person re-identification · Text detection
End-to-end text recognition · Image retrieval · Video surveillance

1 Introduction

Recently, person re-identification (Re-ID) has received considerable attention in computer vision research community. The mainstream state-of-the-art methods for re-identifying person can be categorized into three classes: (1) Feature representation based methods, which extract some handcraft visual features (e.g., color, texture, shape) to define a distinctive descriptor to recognition. Some representative works are [7,20]. (2) Metric learning based methods, which shift the focus from feature selection based efforts to improve Re-ID by learning appropriate distance metrics in the sense of maximizing matching accuracy. These methods are also considered as relative ranking and manifold-based affinity learning problems in many works [1,14]. (3) Deep learning based methods. With the success of deep convolutional neural networks (CNN) in many vision problems,

Kamlesh and P. Xu have contributed equally to this work.

J. Yang et al. (Eds.): CCCV 2017, Part III, CCIS 773, pp. 363–374, 2017.
https://doi.org/10.1007/978-981-10-7305-2_32

several tentative works using CNN for person Re-ID have been proposed recently onto the task to learn to tell different persons since video-based re-id has a large data volume, e.g.,[3,17].

Different from those mainstream Re-ID methods, we notice that under some circumstances every person is associated with a unique text label (e.g., racing bib number). This text label can be considered as an important and effective cue to identify a person. Thus the person Re-ID problem can be turned into a text detection and recognition problem, which is much easier thanks to a great performance of current end-to-end text recognition in the wild.

In this paper, we propose to identify a person by an end-to-end recognition on text that could differentiate one person from others. The main difference between traditional person Re-ID task and our person identification task is depicted in Fig. 1.

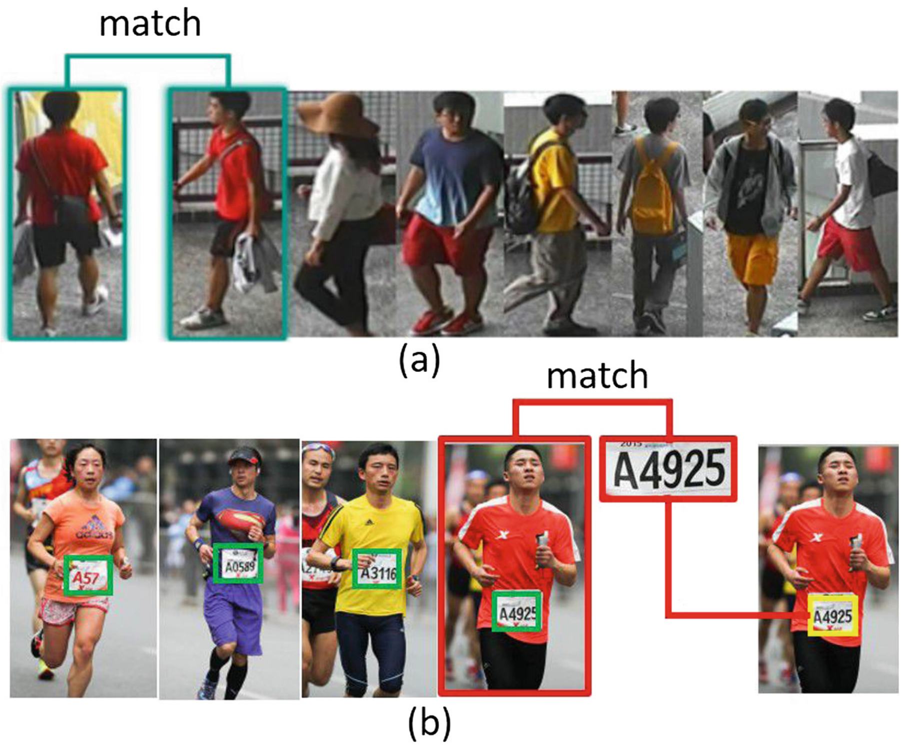

Fig. 1. Comparison of classical person re-identification in (a) and our task in (b), where we aim to identify specific person using end-to-end scene text recognition method.

The proposed idea of using text recognition system to achieve person Re-ID is extremely useful in races such as marathons. Recently, marathons are organized throughout the world and attract many broadcasters and individuals, which results in a large volume of photograph and video data. It is difficult or even impossible to manually identify any particular participant based only on the visual appearance. So each participant is given a unique Racing Bib Number (RBN) designating that contestant. Bib numbers are often printed on a piece of cloth or paper and attached to the body part (e.g., chest) of contestants. We propose to identify a specific person via the text information in terms of RBN recognized using an end-to-end scene text recognition system. This would make it much easier to identify a specific person, search the photo shot of a queried RBN from a large dataset of photos or video recording competition results.

More specifically, since RBN is usually presented as horizontal text attached to the chest of participants, we propose to adopt a state-of-the-art horizontal scene text detector TextBoxes [11] to spot RBN's location. Contrary to most conventional text detectors requiring multiple post-processing steps, TextBoxes is an end-to-end trainable and detects efficiently and accurately horizontal text from scene images. The detection step is followed by an efficient text recognition algorithm called CRNN [15] to transcript detected text. CRNN is also end-to-end trainable and performs well in recognizing scene text. Since RBN usually has a clean background, hopefully, the pipeline of combining TextBoxes and CRNN would recognize correctly RBN from the input image.

The main contributions of this paper are three folds:

(1) We first proposed an alternative and novel idea for person Re-ID via a sophisticated horizontal end-to-end scene text recognition. We proposed an appropriate matching criterion to adapt the text recognition task to person identification.
(2) We collected 9706 marathon images to establish a large and reliable benchmark dataset for person Re-ID task using text cues. To the best of our knowledge, there is no public dataset dedicated for this topic. This dataset will be soon made publicly available.
(3) We conducted a number of experiments confirming the usefulness of the proposed idea which identifies person with end-to-end scene text recognition. The experimental results show that the proposed idea has a considerable potential in person identification, pedestrian search, and video surveillance.

The rest of the paper is organized as follows. Some related work is provided in Sect. 2, followed by our method in Sect. 3. We present in Sect. 4 some experimental results. Finally, we conclude in Sect. 5.

2 Related Work

Person Re-ID is a fundamental task and also a difficult problem in computer vision. This is because the same person may look quite different in different images and two persons may look very similar when exposed to complex backgrounds, viewpoint variations, and so on. Many works about this topic focus on extracting representative visual features for images and/or designing a robust distance metric to compare those features.

In this paper, we focus on identifying participants from Marathon race images. Since each person is associated with a unique RBN, in this case, the person Re-ID can be turned into a text detection and recognition problem. We shortly review some text detection and recognition works in the following. More details can be found in these two comprehensive review papers.

2.1 Text Detection

Text detection aims to localize text regions in images, often in the form of word bounding boxes. Many works have been proposed to solve this issue. They can be divided into three categories: texture based methods [6,10,21], connected-component based methods [4,5,13], and deep learning methods [9,11,19]. Recently, deep learning based methods have been the mainstream and achieved great improvements. In this paper, we adopt TextBoxes [11], and end-to-end trainable method [11] which achieves high performance by applying SSD [12] to text detection. Compared to other deep learning based methods, TextBoxes based on SSD framework uses a fully convolutional pipeline and non-maximum suppression as post processing without extracting regions of interest (ROI). TextBoxes (see [11] for more details) has advantages in the following aspects:

(1) *Hierarchical feature maps for detection*: TextBoxes uses feature map outputs of 6 convolutional layers to detect text of multiple scales. The input image of TextBoxes is of size $300 * 300$, and the sizes of extracted feature maps are $39 * 39$, $20 * 20$, $10 * 10$, $5 * 5$, $3 * 3$, $1 * 1$ respectively.
(2) *Fully convolutional kernel for prediction*: TextBoxes substitutes fully-connected layers with fully-convolutional layers as predictor, which largely reduces computation burden. Different from the $3 * 3$ convolutional predictor used in SSD, TextBoxes uses $1 * 5$ convolution kernels as the predictor, which better fits long text lines.
(3) *Multi-scale aspect ratios for default boxes*: TextBoxes adapts SSD framework by defining 6 aspect ratios for default boxes, including 1, 2, 3, 5, 7, and 10, which better caters to long text lines in natural images.

2.2 Text Recognition

Text recognition aims to translate text regions into human or machine readable character sequences. The text recognition methods can be roughly grouped into two categories: character-based [2,16,18] and word-based recognition [9,15]. In this paper, we adapt a state-of-the-art word-based method called CRNN [15]. CRNN is an end-to-end trainable neural network for image-based sequence recognition, which takes an image as input and outputs recognized string. Its architecture consists of three parts: (1) Convolutional layers which extract a feature sequence from the input image; (2) Recurrent layers which predict a label distribution for each frame; (3) Transcription layer which translates the per-frame predictions into the final label sequence. CRNN (see [15] for more details) shows its superiority in sequence recognition as follows:

(1) It can be directly learned from sequence labels (e.g., words), requiring no detailed annotations (e.g., characters);
(2) It requires neither handcraft features nor preprocessing steps, including binarization or segmentation, component localization, etc.
(3) It is unconstrained to the lengths of sequence-like objects, requiring only height normalization in both training and testing phases.

3 Method

3.1 Text Detection of RBN

Compared to scene text, text of racing bib numbers has the following differences:

(1) Scene text in the wild may appear in the form of a single word, a long text line, or curved lines, which shows large variations both in size and aspect ratio. However, although text of RBN may vary in size, its aspect ratio does not change significantly. This is because RBN is custom made to fixed width and height for a competition.
(2) Scene text in the wild often undergoes complex background and noises. Unlikely, RBN has distinguished appearance between black text and pure color background. It is easier to detect RBN text than scene text in the wild.

Based on these two differences, we propose to adapt TextBoxes [11] and make the following changes to detect text of RBN:

(1) Original TextBoxes uses 6 aspect ratios for default boxes (see Sect. 2.1). In the case of RBN text detection, almost all the RBN texts have an aspect ratio lower than 5 even for twisted and deformed cases. In fact, based on our observation, the aspect ratio for RBN text is near 3. So we only reserve aspect ratios 1, 2, and 3 of TextBoxes shortly reviewed in Sect. 2.1. In this way, we can achieve higher accuracy and meanwhile remove unnecessary computations.
(2) Since the size of images in our dataset is 1200 × 800, resizing them to 300*300 may fail to capture RBN of small size due to low resolution. So we alternately resize input images to 300 × 300 and 700 × 700 every 200 batches. Hopefully, higher resolution of inputs helps to detect small texts explicitly.
(3) Finally, original TextBoxes uses 1 × 5 convolutional kernels as predictor of score and offsets for scene text detection in the wild. In the case of RBN text, we do not need such a long kernel. Instead, we use 1 × 3 convolutional kernel.

The architecture of our adapted TextBoxes is illustrated in Fig. 2.

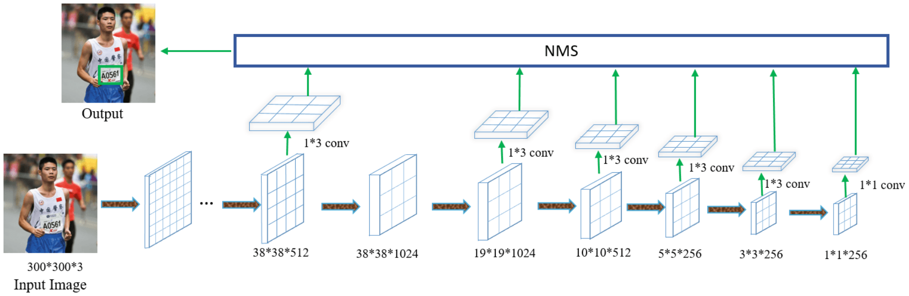

Fig. 2. Architecture of adapted TextBoxes [11] for RBN text detection.

3.2 Text Recognition of RBN

We adopt CRNN [15] shortly reviewed in Sect. 2.2 to recognize text bounding boxes produced by modified TextBoxes detailed in Sect. 3.1. Since RBN text is mainly composed of pure numbers or an English character followed by several numbers, which is a special case of general scene text concerned in original CRNN. For scene text recognition, original CRNN uses the synthetic dataset (Synth, 800K) released by Jaderberg *et al.* [8] as the training data and tests on all other real-world test datasets without any fine-tuning on their training data. The images in Synth dataset have a distinguished dictionary, which mainly consists of English characters, thus very likely to work poorly on our task if using directly the released trained model. Consequently, we propose to use an alternative synthetic dataset to retrain the model of CRNN. With the dictionary of our marathon dataset, we employ the method in [8] to synthesize 100K images looking authentic compared to marathon dataset. We retrain the pretrained model released by CRNN with such generated 100K synthetic images and then fine-tune the model with our train data.

4 Experiments

4.1 Dataset Creation

We have collected by ourselves 9706 images from different events of marathon under varying imaging environments. We divide this dataset into two subsets: training set containing 8706 images and testing set with the other 1000 images. The collected dataset exhibits following multiple challenges: multiple candidates of different scales, deformed text with low illumination, and text in blur or in low resolution. For all images, text regions are manually labeled with their bounding boxes ($x1$, $y1$, $x2$, $y2$, $x3$, $y3$, $x4$, $y4$), corresponding strings (s), and an indicator (dif) denotes whether the underlying text is difficult to recognize. This amounts up to a vector of 10 entries. The dictionary of the annotations consists mainly of numbers and a few English characters.

We believe our collected dataset can serve as a standard benchmark for end-to-end text recognition as well as for person re-identification where a unique bib number identifies each individual runner.

4.2 Evaluation Protocols

Detection Measurement. We apply classical and popular precision, recall, and F-measure metrics to assess text detection performance. A detection box *dt* is considered to match a ground truth box *gt* if the intersection-over-union IoU exceeds a given threshold (set to 0.5 in all our experiments). The IoU is defined as $IoU = \frac{|gt \bigcap dt|}{|gt \bigcup dt|}$, where $|\cdot|$ stands for the cardinality. When more than one detection is matched to the same ground truth, only the detection with the maximum IoU is kept, the rest are matched to null (i.e., unmatched). This means we stick to one-to-one match strategy. Then the precision P, recall R, and $F_{measure}$ are computed by

$$P = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$F_{measure} = \frac{2 * P * R}{P + R} \tag{3}$$

where TP, FP, and FN are the number of hit detection boxes, incorrectly identified detection boxes (unmatched detections), and missing ground truth boxes(unmatched ground truth) respectively.

End-to-End Recognition Measurement. We use edit distance *ed* as evaluation metric for recognition. Following the same matching strategy as detection process, we calculate the edit distances between all matching pairs. If a detection is matched to null, then an empty string will be taken as the ground truth text. Similarly, if a ground truth is matched to null, we also calculate its edit distance to empty string. The edit distances are summarized and divided by the number of test images. The resulting average edit distance AED is given by:

$$AED = \frac{\sum_{i=1}^{N_t} \sum_{j=1}^{N_m} ed(dt, gt))}{N_t}, \tag{4}$$

where N_t is the number of test images, N_m is the number of matched pairs. This average edit distance AED is considered as the metric. A smaller AED means a better performance.

Person Identification Measurement. AED is an appropriate measurement for overall end-to-end recognition, but not adapted to measure the accuracy of person identification. Because AED calculates the edit distance between the recognized sequence and the corresponding ground truth sequence. Whereas, in the case of person Re-ID via RBN recognition, only when all characters of the recognized text are completely the same as the RBN, the identification can be deemed as finding correctly the queried identity. So we introduce a more precise measurement of matching criterion for person identification. Let hit denote whether we identify the right person corresponding to the queried RBN or not. If the ed of a recognized string is 0, hit is equal to 1, otherwise 0. We use ID_Fm, ID_P, ID_R defined in the following to measure the identification performance on the overall test dataset:

$$ID\_P = \frac{\sum_{i=1}^{N_t} \sum_{j=1}^{N_r} hit}{\sum_{i=1}^{N_t} N_r} \tag{5}$$

$$ID\_R = \frac{\sum_{i=1}^{N_t} \sum_{j=1}^{N_r} hit}{\sum_{i=1}^{N_t} N_g} \tag{6}$$

$$ID\_Fm = 2 * \frac{ID\_P * ID\_R}{ID\_P + ID\_R} \tag{7}$$

where N_r is the number of extracted boxes to be recognized on a single test image, N_g denotes the number of ground truth boxes correspondingly, ID_P, ID_R, ID_Fm for precision, recall, and F_measure of identification respectively. A larger ID_Fm means a better performance on person identification through end-to-end text recognition.

4.3 Implementation Details

RBN Detection. We conduct two comparison experiments in detection experiment depending on whether to use synthetic detection data to pretrain or not:

(1) For experiment using synthetic data, the model is trained for 500k iterations using a learning rate 0.001 for the first 40k iterations, and 0.0001 for the rest iterations. The batch size for SGD is 32 for 300 * 300 input. Then the model is fine-tuned on real marathon data for 14.1k iterations with the learning rate set to 0.0001. The batch size is set to 8 for resized images into 300 * 300 or 700 * 700 alternatively.

(2) For experiment without using synthetic data, we consider it as a raw training on pretrained vgg-16 model. The model is trained for 23.5k iterations with the learning rate set to 0.001 for the first 3.5k iterations, and 0.0001 for the rest 20k iterations. The batch size for SGD is set to 8.

RBN End-to-End Recognition. We first use 100k synthetic images specifically synthesized for marathon dataset to train on the released pretrained model by CRNN for 50k iterations. Then we use real data to train for 330k iterations. The two trainings are both optimized by ADADELTA to automatically calculate per-dimension learning rates.

Experimental Environment. The text detection method is implemented using Caffe, and the recognition method is implemented using Torch. All the experiments are conducted on a workstation with a 1.9 GHz Intel(R) Xeon(R) E5-2609 CPU, 64 GB RAM and a NVIDIA GTX TitanX GPU.

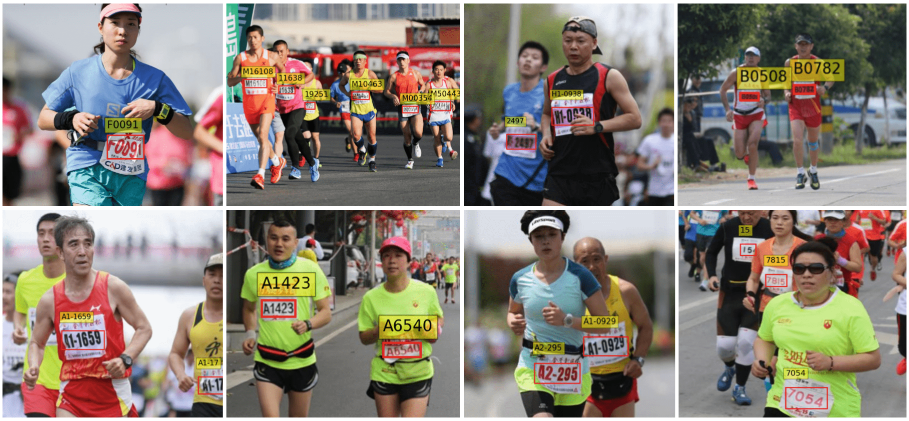

Fig. 3. Performance of our proposed method to identify specific person using an end-to-end scene text recognition method.

4.4 Results

Since there is no previous work on the person Re-ID task using an end-to-end scene text recognition method, we evaluate our proposed Textboxes + CRNN pipeline on self-collected marathon dataset, and compare it to another component based detection pipeline FCN + CRNN. Since they both use the same recognition model CRNN, we compare both the detection and recognition performance.

Some quantitative results are depicted in Table 1. It can be seen that our proposed method achieves remarkable performance, which is better than the other baseline in both detection and recognition performance. This improvement is consistent with the detection performance gap between TextBoxes and FCN. And the ID_Fm of the best model TCRM (TextBoxes + CRNN 0+ Raw + Multi) can reach above 0.7, nearly 1% superior to FCN + CRNN pipeline. Furthermore, our end-to-end text recognition method is very fast, the detection phase takes 0.076 s for testing in single scale (0.19 s for testing in multi scale), and the recognition phase takes 0.95 s for testing with a lexicon.

As demonstrated by these results, an end-to-end scene text recognition can also be applied to person identification and achieve remarkable performance.

Table 1. The performance on the marathon dataset. The sign in method name TC for TextBoxes + CRNN, R/S for raw finetuning on vgg16 model or train first with synthetic data on vgg16 and then finetune, M/S for testing in multi/single scale. The compare method in the table is tested by FCN + CRNN. AED (resp. AED_LEX) for average edit distance without (resp. with) a lexicon, and ID_Fm (resp. ID_Fm_LEX) for Fmeasure without (resp. with) lexicon.

Method	P	R	F_measure	AED	AED_LEX	ID_fm	ID_fm_LEX
TCRM	**0.9294**	0.8931	**0.9109**	**1.529**	**3.389**	**0.6167**	**0.7054**
TCSM	0.9156	**0.8993**	0.9074	1.638	3.463	0.6038	0.6989
TCRS	0.9197	0.8785	0.8986	1.662	3.526	0.5978	0.6929
TCSS	0.9226	0.8764	0.8989	1.718	3.646	0.5957	0.6909
FCN + CRNN	0.9026	0.8767	0.8895	1.691	3.577	0.5954	0.6957

Some qualitative illustrations of the best identification performances are shown in Fig. 3. We can see that the proposed end-to-end scene text recognition handles well many different circumstances, such as text regions of different scales, unbalanced illumination, low resolution, deformed text, block or blur, etc. Furthermore, we can see that some recognition result can surpass human recognition. For many text regions, the recognition can even learn more than human can annotate. For those person whose text regions are blocked or with incomplete annotation, we can rank the recognition results according to edit distance and give top 10 results. Towards these candidates of most possibilities, we can further involve standard person re-identification methods to identify persons precisely.

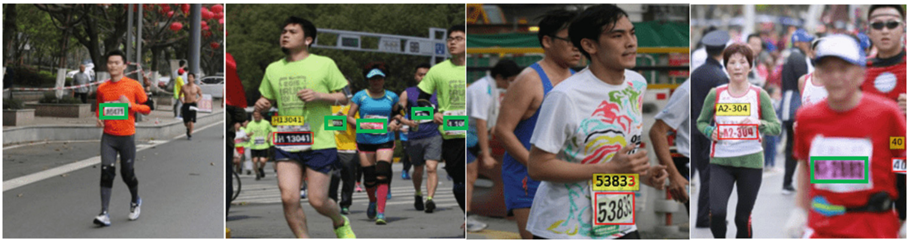

Fig. 4. Some failure cases of our proposed method. Red boxes and black annotations are correct end-to-end detection. Green Boxes represent missing boxes. Red annotations stands for wrong recognized characters. (Color figure online)

Some failure cases are also shown in Fig. 4, where we miss a part of boxes or falsely recognize texts. The proposed method does not handle well small text in blur and low resolution, unbalanced illumination. Thus the proposed person identification is still far from ideal and has lots of potentials to improve. One perspective is to integrate feature representation of image or metric learning method in the proposed method to improve the performance for person identification.

5 Conclusion

We have presented in this paper a novel alternative method for person identification based on current sophisticated end-to-end scene text recognition system. We adopted TextBoxes and CRNN to accomplish this task. Moreover, we have collected a large dataset of marathon images for benchmarking the proposed method. This dataset will be soon made publicly available for combination of text recognition and person identification. Furthermore, we proposed an appropriate measurement for evaluation identification performance and compared with another pipeline to confirm our proposal. Experimental results demonstrate promising results of the proposed method, and a high potential to explore in person Re-ID, video surveillance, and image retrieval. In the future, we plan to integrate some traditional person Re-ID methods into our proposed pipeline to further improve identification performance on our benchmark and make more comprehensive comparisons.

Acknowledgements. This work was supported by National Natural Science Foundation of China 61703171 and 61222308.

References

1. Bai, S., Bai, X., Tian, Q.: Scalable person re-identification on supervised smoothed manifold. CoRR abs/1703.08359 (2017)
2. Bissacco, A., Cummins, M., Netzer, Y., Neven, H.: PhotoOCR: Reading text in uncontrolled conditions. In: Proceedings of ICCV (2013)
3. Bromley, J., Bentz, J.W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E., Shah, R.: Signature verification using a "siamese" time delay neural network. IJPRAI **7**(4), 669–688 (1993)
4. Busta, M., Neumann, L., Matas, J.: Fastext: efficient unconstrained scene text detector. In: Proceedings of ICCV (2015)
5. Epshtein, B., Ofek, E., Wexler, Y.: Detecting text in natural scenes with stroke width transform. In: Proceedings of CVPR (2010)
6. Goto, H., Tanaka, M.: Text-tracking wearable camera system for the blind. In: 10th International Conference on Document Analysis and Recognition (2009)
7. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5302, pp. 262–275. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88682-2_21
8. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Synthetic data and artificial neural networks for natural scene text recognition. CoRR abs/1406.2227 (2014)
9. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Reading text in the wild with convolutional neural networks. IJCV **116**(1), 1–20 (2016)
10. Li, H., Doermann, D.S., Kia, O.E.: Automatic text detection and tracking in digital video. IEEE Trans. Image Process. **9**(1), 147–156 (2000)
11. Liao, M., Shi, B., Bai, X., Wang, X., Liu, W.: TextBoxes: a fast text detector with a single deep neural network. In: Proceedings of AAAI (2017)

12. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
13. Neumann, L., Matas, J.: A method for text localization and recognition in real-world images. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010. LNCS, vol. 6494, pp. 770–783. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-19318-7_60
14. Prosser, B.J., Zheng, W., Gong, S., Xiang, T.: Person re-identification by support vector ranking. In: BMVC 2010 (2010)
15. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. TPAMI **39**(11), 2298–2304 (2016)
16. Wang, T., Wu, D.J., Coates, A., Ng, A.Y.: End-to-end text recognition with convolutional neural networks. In: Proceedings of ICPR (2012)
17. Wu, L., Shen, C., van den Hengel, A.: PersonNet: person re-identification with deep convolutional neural networks. CoRR abs/1601.07255 (2016)
18. Yao, C., Bai, X., Shi, B., Liu, W.: Strokelets: A learned multi-scale representation for scene text recognition. In: Proceedings of CVPR (2014)
19. Zhang, Z., Zhang, C., Shen, W., Yao, C., Liu, W., Bai, X.: Multi-oriented text detection with fully convolutional networks. In: Proceedings of CVPR (2016)
20. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: Proceedings of CVPR (2013)
21. Zhong, Y., Karu, K., Jain, A.K.: Locating text in complex color images. Pattern Recogn. **28**(10), 1523–1535 (1995)