Proceedings of the 2016 IEEE
International Conference on Robotics and Biomimetics
Qingdao, China, December 3-7, 2016

STD: A Stereo Tracking Dataset for Evaluating Binocular Tracking Algorithms

Zheng Zhu, Wei Zou, Qingbin Wang, and Feng Zhang

Abstract—**In this paper, a Stereo Tracking Dataset is proposed for evaluating binocular tracking algorithms. The dataset contains stereoscopic videos which are collected by our mobile platform in different scenarios and videos that are available publicly. All sequences are carefully synchronized and rectified, and the ground truth of object is annotated by authors. Both raw and processed sequences are provided in the dataset. We also develop a Scalable and Occlusion-aware Multi-cues Correlation Filter Tracker (SOMCFT) and evaluate it on the STD. The SOMCFT framework fuses different clues in confidence map level and uses depth information to handle scale changes and occlusion. Quantitative evaluation on STD demonstrates effectiveness of the proposed dataset. All data, including stereo image pairs, calibrations, annotations and attributes, are available for research purposes and comparative evaluation on https://github.com/zhengzhugithub/StereoTracking.**

I. Introduction

Visual object tracking always plays an important role in computer vision with the application in automatic driving [1], human-machine interactions [2] and robot perception [3]. The core problem of tracking is to detect and locate the object with appearance variations in the changing scenario [4]. Besides, tracking is a time-critical problem. These two aspects, robustness and efficiency, are main development directions of the recent tracking approaches.

Monocular image-based tracking methods are easily corrupted by the various noises and can't handle occlusions effectively [19, 20, 21, 22, 23, 30, 31, 32, 33, 34, 35]. Depth information can be adopted to alleviate these problems. In recent years, popularity of affordable depth sensors such as Microsoft Kinect and Asus Xtion make depth acquisition easy, thus booming so-called RGB-D tracking algorithms [29, 18]. Another common method to obtain depth information is using binocular cameras. The stereo setups can perform efficiently in outdoor environment and can overcome the distance limit. Almost all of primates have binocular systems and many robot systems are also equipped with binocular vision to tracking targets [7, 8, 12, 13, 14]. These stereo vision systems exploit the additional information obtained by exploiting the stereo geometry, namely the depth information. This extra depth information can be a useful cue for visual tracking with handling scale variations and occlusions. Besides, depth information can be used as a feature to discriminate the target from the background. So, stereo vision structure can boost the robustness of the visual object tracking.

For stereo tracking and tracking algorithms with depth information, there are some datasets to evaluate. The Princeton Datasets [29] and the BoBoT-D Benchmark [18] are both recorded with Microsoft Kinect. The former contains 5 validation and 95 test sequences while the latter consists of 5 sequences with object rotation, occlusion and scale changes. The raw dataset in KITTI Benchmark [5] provides synced and rectified color stereo sequences that are recorded with 2 color cameras equipped on a VW Passat station wagon. The object tracking dataset in this benchmark consists of 21 training sequences and 29 test sequences, but the bounding boxes are not proper for evaluating a single target tracker. What is more, the dataset is not categorized with attributes, making it difficult to evaluate certain characteristic of stereo tracking algorithms. The New College Dataset [28] contains grayscale stereo imagery without annotated bounding boxes. The Malaga Dataset [27] is gathered entirely in urban scenarios with a car equipped with stereo camera and also without bounding boxes. These stereo datasets are not suitable for evaluating stereo since most objects disappear in the view and never appear again. Besides, most scenes are not annotated with bounding box or the annotation is not suitable for tracking, which makes it difficult for comprehensive performance evaluation of stereo trackers. So we collect and pre-process stereoscopic videos, which are publicly available on the internet to develop a Stereo Tracking Dataset (STD). And we also collect and annotate stereo image pairs using our mobile robot in different scenarios. Table 1 lists an overview of above datasets. As shown in Table 1, STD is a binocular dataset which contains lab/outdoor scenarios and are fully annotated. Besides, the sequences in STD are labeled with attributes, which makes it convenient for evaluating a specific attribute of a tracker. To the best of our knowledge, this is the first complete datasets for evaluating stereo tracking algorithm.

This study was supported by the National High Technology Research and Development Program of China ("863 Program") under Grant No. 2015AA042307, and in part by the National Natural Science Foundation of China under NO. 61573347 and 61403378.

The authors are with Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (email: {zhuzheng2014, wei.zou, wangqingbin2012, feng.zhang }@ia.ac.cn). Zheng Zhu and Qingbin Wang are also with University of Chinese Academy of Science, Beijing, 100190, China.

978-1-5090-4364-4/16/$31.00 ©2016 IEEE

Table 1 Comparison of STD and other tracking datasets

Name	Type	Scenarios	Annotation	Attributes	Suitable for evaluating stereo trackers
STD (proposed)	binocular	lab/outdoor	yes	yes	yes
CVPR Dataset [24]	monocular	lab/outdoor	yes	yes	no
VOT Challenges[25,17]	monocular	lab/outdoor	yes	no	no
Princeton Dataset [29]	RGB-D	almost in lab	partial	no	no
BoBoT-D Dataset [18]	RGB-D	lab	yes	no	no
KITTI Dataset [5]	binocular	outdoor	partial	no	no
New College Dataset [28]	binocular	outdoor	no	no	no
Malaga Dataset [27]	binocular	outdoor	no	no	no

II. Stereo Tracking Dataset

A. Capturing and Pre-processing

Figure 1 shows our mobile robot platform used in capturing STD. The computer and a pair of cameras are mounted on a wheel platform. The intrinsic parameters, distortion and extrinsic parameters are calibrated at first. The intrinsic parameters matrices of cameras could be calibrated using Zhang's calibration method as [15]:

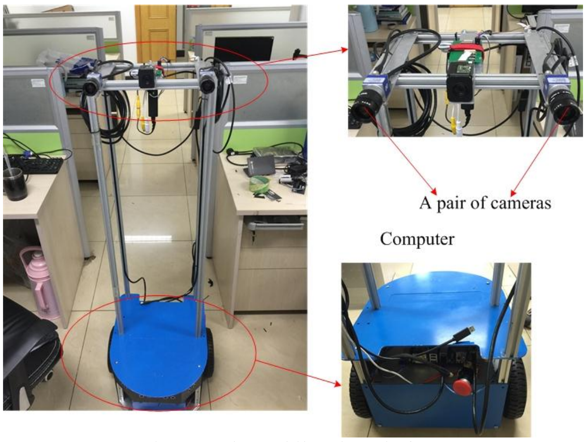

Figure 1 The mobile robot platform

$$M_{in} = \begin{pmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{pmatrix} \tag{1}$$

where (f_x, f_y) are magnification factors, (u_0, v_0) are image coordinates of the primary points of cameras.

The distortion parameter matrices of the left and right cameras could be calibrated using Brown's calibration method [16]:

$$K_d = (k_1, k_2, p_1, p_2, k_3) \tag{2}$$

where k_1, k_2, and k_3 are radial distortion parameters, p_1, p_2 are tangential distortion parameters.

The binocular extrinsic parameters could be calibrated using Zhang's stereo calibration method [15]:

$$R = \begin{pmatrix} n_x & o_x & a_x \\ n_y & o_y & a_y \\ n_z & o_z & a_z \end{pmatrix} \quad p = \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix} \tag{3}$$

where R is 3×3 rotation matrix and p is 3×1 translation vectors.

After calibrations, the depth map can be obtained from binocular images. The rectified images are obtained using the above calibrating parameters.

B. Details of sequences

The STD contains 23 sequences which are annotated in each frame. The sequences *jiagang2*, *jiagang3*, *MI1*, *MI2*, *talk*, *rotation2*, *rotation1*, *zhengzhu4*, *zhengzhu3*, *zhengzhu2*, *zhengzhu1*, *tengzhang2*, and *tengzhang1* are recorded by our mobile robots equipping a pair of cameras. The scenarios are mainly in laboratory and the tracked objects are mainly faces or heads. Since our mobile robot is inconvenient to work outdoor at present, we collect 10 sequences from public datasets [6, 5] which are recording in outdoor scenarios. These sequences are *BAHNHOF*, *cyclist*, *LINTHESCHER*, *LOEWENPLATZ*, *person2*, *LOEWENPLATZ2*, *PEDCROSS2*, *person3*, *SUNNYDAY*, and *person1*. Note that these 10 sequences are not annotated and not suitable for evaluating tracking algorithms. So we pre-process and annotate them to make them compatible with the format of STD. Fig.2 gives the screenshot of first frame of each sequence and corresponding bounding boxes.

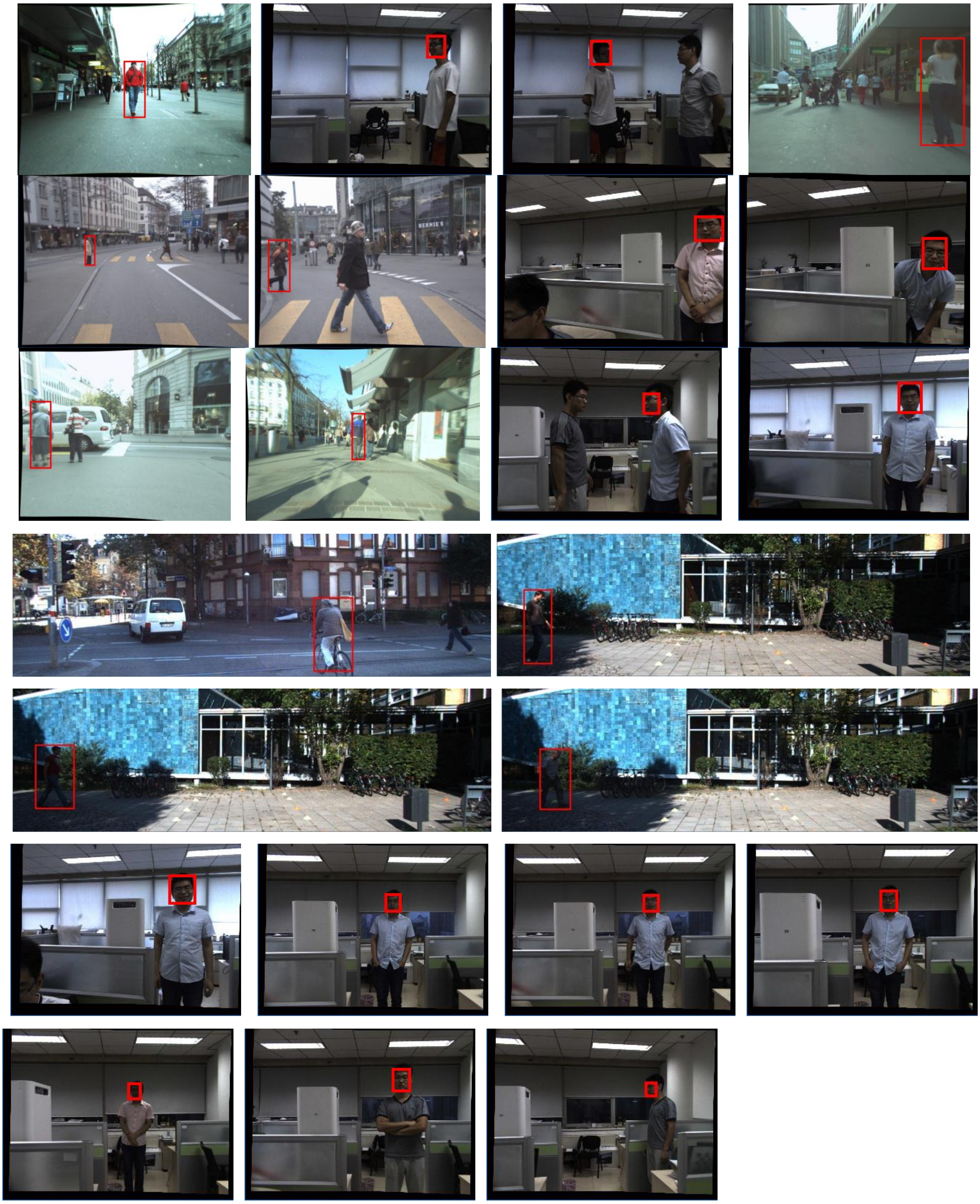

Figure 2 screenshot of first frame of each sequence and corresponding bounding boxes, they are ordered with *BAHNHOF*,*jiagang2*,*jiagang3*, *LINTHESCHER*, *LOEWENPLATZ*, *LOEWENPLATZ2*, *MI1*, *MI2*, *PEDCROSS 2*, *SUNNY DAY*, *talk rotation2*, *cyclist*, *person1*, *person2*, *person3*, *rotation1*, *zhengzhu4*, *zhengzhu3*, *zhengzhu2*, *zhengzhu1*, *tengzhang2*, *tengzhang1*.

C. *Dataset with attributes*

To better evaluate and analyze the strength and weakness of the tracking approaches, the videos are categorized with 6 attributes based on different challenging factors including occlusion (OCC), scaling (SCA), illumination variation (IV), background cluttering (BC), pose variation (PV), cameras motion (CM), which are summarized in Table 2. √ in blank means specific attributes for the videos. What is more, √* means severe challenges (main difficulties) in videos, such as fully occlusion, severe illumination changes, large-angle pose variation and fast cameras motion. The last row of table gives the number of sequences with the specific attributes.

Table 2 Sequences with attributes, modest and severe challenges.

sequences	OCC	SCA	IV	BC	PV	CM
MI2	√*			√	√	
MI1	√*		√	√	√	
talk		√		√*	√	
tengzhang1	√	√	√		√*	
zhengzhu1		√*	√		√*	
zhengzhu2		√	√*		√	
tengzhang2	√	√*			√*	√
zhengzhu3	√				√	√
zhengzhu4		√	√		√	√
BAHNHOF		√*				√*
LINTHESCHER	√*	√*	√	√*	√	√*
PEDCROSS 2	√*	√*	√	√		√*
SUNNY DAY	√	√	√*	√*		√*
LOEWENPLATZ		√*	√		√	
LOEWENPLATZ2				√	√*	
cyclist		√*	√*		√	√*
person1	√		√*		√	
person2			√*		√*	
person3	√		√*		√	
Jiagang2	√		√*		√*	
Jiagang3				√*	√	
Rotation1	√*				√*	
Rotation2	√*				√*	
Number Total/ severe	13/6	12/7	14/7	8/4	20/8	8/5

III. Experiments

A. *Scalable and occlusion handling Multi-cues Correlation Filter Tracker*

The Scalable and Occlusion-aware Multi-cues Correlation Filter Tracking (SOMCFT) framework is described in *Algorithm 1* to demonstrate the usefulness of STD.

```
Algorithm 1: SOMCFT algorithm
Input: grayscale image $I_g$ and depth image $I_d$
Output: the position of tracked target
for first frame to last frame do
  if first frame do
    Initial the tracker
    segmenting the target region $R^1_{target}$
    set initial depth value $d^1_{target}$ with mean depth of $R^1_{target}$
  else
    multi-cue correlation filters tracking
    scale and occlusion handling processing.
  end if
end for
Return the position of tracked target
```

B. Evaluation Methodology

The best methodology to evaluate trackers is still a debatable subject. Recently, researchers argue for the use of precision plot [11] and success plot [24]. The precision plot shows, for a range of distance thresholds, the percentage of frames that the tracker is within that distance of the ground truth. The score when the threshold is 20 pixels can be regarded as the representative precision score. This precision score is written as PS20 in the following aspects. The success plot shows the radios of successful frames at the thresholds varied from 0 to 1. A successful frame is counted when its overlap is larger than the given threshold. The area under curve (AUC) of each success plot is used to rank the tracking algorithm.

In our experiments, both precision plot and success plot are adopted to evaluate the performance of trackers. Besides, PS20 and AUC are also reported for quantitative evaluation. Firstly, the proposed SOMCFT tracker is compared with state-of-the-art RGB trackers, namely TLD [10], STRUCK [9], KCF [21] and DSST [23]. Then the proposed tracking algorithm is evaluated with trackers which utilize depth information since it is used in our tracker. The performance of two depth tracker, SAMF+Depth [22] and RGBDOcc+OF [29] are compared in experiments.

C. Overall performance

Fig.4 shows the overall performance of the evaluated tracking algorithms in terms of precision and success plots. Left is the performance score of success plot while right is performance score of precession plot. The proposed SOMCFT tracking algorithm ranks first both in success and precision score. In the success plot, the AUC score of SOMCFT is 0.5516 which outperforms the second OccOF method by 5.4%. Meanwhile, in the precision plot, the proposed SOMCFT algorithm achieves the score of 0.8182, which outperforms the second OccOF method by 14.69%. Since the proposed tracker adopt multi-cue integration correlation, it is robust to changes such as illumination and pose variation, and can achieve competitive performance to the other trackers. Furthermore, the SOMCFT tracking algorithm is scalable and occlusion-aware, all these strategy guarantee the tracker with the best performance.

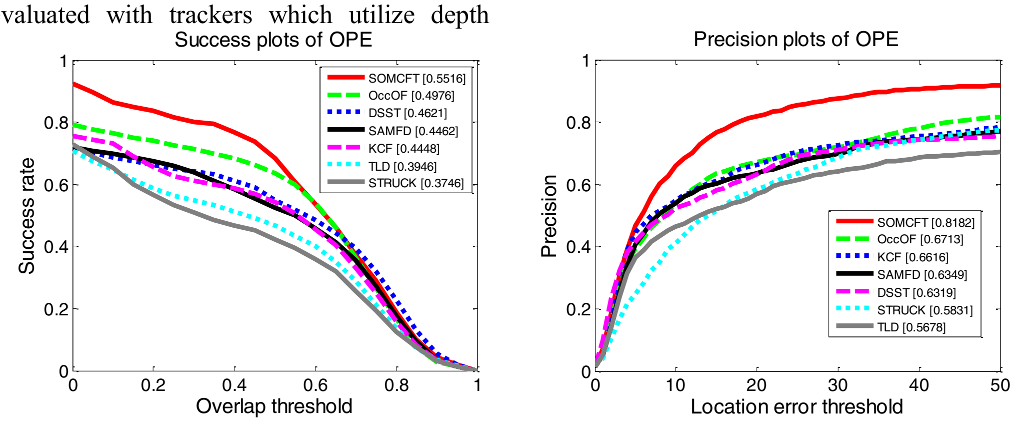

Figure. 4: The success plots and precision plots for the 7 trackers. The performance score of success plot is the AUC value while the performance score for each tracker is shown in the legend. The performance score of precession plot is at error threshold of 20 pixels. Best viewed on color display.

D. Attribute-based performance

To better evaluate and analyze the strength and weakness of the tracking approaches, the trackers are evaluated with 6 attributes. The tracking results in terms of success and precision score are listed in Table 3 and Table 4. The red fonts indicate the best performance, the blue fonts indicate the second best ones, and the green fonts indicate the third best ones. As shown in Table 3, the proposed SOMCFT tracker ranks first in precision score of overall performance and 4 attribute subsets. It is shown in Table 4 that SOMCFT ranks first in success score of overall performance and 4 attribute subsets.

TABLE 3: Score of success plot (best viewed on a color display).

attribute	SOMCFT	TLD[10]	STRUCK[9]	KCF[21]	DSST[23]	Occ+OF[29]	SAMFD[22]
OCC	0.5044	0.3747	0.4060	0.4093	0.4060	0.4429	0.3897
SCA	0.5113	0.4376	0.2965	0.4735	0.5103	0.4838	0.5211
IV	0.4824	0.2760	0.3050	0.3881	0.3704	0.4261	0.3792
BC	0.5340	0.3859	0.4595	0.4365	0.5192	0.4766	0.5620
PV	0.5733	0.3762	0.3590	0.4419	0.4242	0.4956	0.4196
CM	0.5440	0.4883	0.3723	0.5692	0.6017	0.5109	0.5594
Overall score	0.5516	0.3946	0.3746	0.4448	0.4621	0.4976	0.4462

TABLE 4: Score of precision plot (best viewed on a color display).

attribute	SOMCFT	TLD[10]	STRUCK[9]	KCF[21]	DSST[23]	Occ+OF[29]	SAMFD[22]
OCC	0.7592	0.5678	0.6239	0.5834	0.5319	0.6131	0.5207
SCA	0.7945	0.6173	0.4897	0.7553	0.6780	0.6200	0.7617
IV	0.7536	0.4364	0.4875	0.6223	0.5109	0.5803	0.5730
BC	0.7351	0.5225	0.6251	0.5948	0.6249	0.5479	0.7234
PV	0.8423	0.5716	0.5592	0.6611	0.6054	0.7092	0.6208
CM	0.7919	0.5918	0.6050	0.8464	0.7151	0.6086	0.7680
Overall score	0.8182	0.5678	0.5831	0.6616	0.6319	0.6713	0.6349

IV. Conclusions

A Stereo Tracking Dataset is proposed for evaluating binocular tracking algorithms in this paper. The dataset contains stereoscopic videos which are collected by our mobile platform in different scenarios and videos that are available publicly. All sequences are carefully synchronized and rectified, and the ground truth of object is annotated by authors. What is more, the sequences in STD are labeled with attributes, which makes it convenient for evaluating a specific attribute of a tracker. Meanwhile, we evaluate the proposed SOMCFT framework on STD to demonstrate the effectiveness of the dataset. In the future work, we intend to build an online website to provide a benchmark for evaluating stereo trackers.

References

[1] Jazayeri A, Cai H, Zheng J Y, et al. Vehicle detection and tracking in car video based on motion model[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2011, 12(2): 583-595.

[2] Shih S W, Liu J. A novel approach to 3-D gaze tracking using stereo cameras [J]. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2004, 34(1): 234-245.

[3] Choi W, Pantofaru C, Savarese S. Detecting and tracking people using an rgb-d camera via multiple detector fusion[C]. *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, 2011: 1076-1083.

[4] Wu Y, Lim J, Yang M H. Online object tracking: A benchmark[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013: 2411-2418.

[5] Geiger A, Lenz P, Stiller C, et al. Vision meets robotics: The KITTI dataset[J]. *The International Journal of Robotics Research*, 2013, 32(11): 1231-1237.

[6] Ess A, Leibe B, Schindler K, et al. Robust multiperson tracking from a mobile platform[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, 31(10): 1831-1846.

[7] Chen Y, Shen Y, Liu X, et al. 3D object tracking via image sets and depth-based occlusion detection[J]. *Signal Processing*, 2015, 112: 146-153.

[8] Zhong B, Shen Y, Chen Y, et al. Online learning 3D context for robust visual tracking[J]. *Neurocomputing,* 2015, 151: 710-718.

[9] Hare S., S Golodetz, A Saffari, V Vineet, M M Cheng, S Hicks, and P Torr. Struck: Structured Output Tracking with Kernels[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.

[10] Kalal Z, Mikolajczyk K, Matas J. Tracking-learning-detection [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, 34(7): 1409-1422.

[11] Babenko B, Yang M H, Belongie S. Robust object tracking with online multiple instance learning [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, 33(8): 1619-1632.

[12] N. Pateromichelakis, A. Mazel, M. A. Hache, T. Koumpogiannis, et. al. Head-eyes system and gaze analysis of the humanoid robot Romeo, in *Proceedings of 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014,1374-1379.

[13] I. Lütkebohle, F. Hegel, S. Schulz, M. Hackel, et. al. The Bielefeld anthropomorphic robot head 'Flobi', in *Proceedings of IEEE International Conference on Robotics and Automation*, 2010, 3384–3391.

[14] T. Kishi, T. Otani, N. Endo, P. Kryczka, et. al. Development of expressive robotic head for bipedal humanoid robot, in *Proceedings of 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, 4584-4589.

[15] Z. Zhang, A flexible new technique for camera calibration, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.22, no.11, pp: 1330-1334, 2000.

[16] D. C. Brown, Close-range camera calibration, *Photogrammetric engineering*, vol.37, no.8, pp: 855-866, 1971.

[17] M. Kristan, R. Pflugfelder, A. Leonardis, J. Matas, L. Cehovin, G. Nebehay, et al, The visual object tracking VOT2014 challenge results[C], *European Conference on Computer Vision*, 2014: 191–217.

[18] G. García, D. Klein, J. Stückler, S. Frintrop, and A. Cremers. Adaptive multi-cue 3D tracking of arbitrary objects. *Pattern Recognition*, pages 357–366. 2012.

[19] Bolme D S, Beveridge J R, Draper B A, et al. Visual object tracking using adaptive correlation filters[C] Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010: 2544-2550.

[20] Henriques J F, Caseiro R, Martins P, et al. Exploiting the circulant structure of tracking-by-detection with kernels [C], *European Conference on Computer Vision*, 2012: 702-715.

[21] Henriques J F, Caseiro R, Martins P, et al. High-speed tracking with kernelized correlation filters [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 37(3): 583-596.

[22] Li Y, Zhu J. A scale adaptive kernel correlation filter tracker with feature integration[C], *European Conference on Computer Vision*, 2014: 254-265.

[23] Danelljan M, Häger G, Khan F, et al. Accurate scale estimation for robust visual tracking[C], *British Machine Vision Conference*, 2014:1-5.

[24] Wu, Yi, Jongwoo Lim, and Ming-Hsuan Yang. "Object tracking benchmark." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.9 (2015): 1834-1848.

[25] Kristan M, Matas J, Leonardis A, et al. The Visual Object Tracking VOT2015 Challenge Results[C] *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2015: 1-23.

[26] Felsberg M, Berg A, Hager G, et al. The thermal infrared visual object tracking VOT-TIR2015 challenge results[C] *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2015: 76-88.

[27] Blanco-Claraco J L, Moreno-Dueñas F Á, González-Jiménez J. The Málaga urban dataset: High-rate stereo and LiDAR in a realistic urban scenario[J]. *The International Journal of Robotics Research*, 2014, 33(2): 207-214.

[28] Smith M, Baldwin I, Churchill W, et al. The new college vision and laser data set[J]. *The International Journal of Robotics Research*, 2009, 28(5): 595-599.

[29] Song S, Xiao J. Tracking revisited using rgbd camera: Unified benchmark and baselines[C] *Proceedings of the IEEE International Conference on Computer Vision*. 2013: 233-240.

[30] Danelljan M, Khan F, Felsberg M, et al. Adaptive color attributes for real-time visual tracking[C] *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014: 1090-1097.

[31] Yoon J H, Kim D Y, Yoon K J. Visual tracking via adaptive tracker selection with multiple features[C]. *European Conference on Computer Vision*, 2012: 28-41.

[32] Lan X, Ma A, Yuen P. Multi-cue visual tracking using robust feature-level fusion based on joint sparse representation[C] *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2014: 1194-1201.

[33] Chen D, Yuan Z, Hua G, et al. Description-discrimination collaborative tracking[C]. *European Conference on Computer Vision*, 2014: 345-360.

[34] Zhong W, Lu H, Yang M H. Robust object tracking via sparsity-based collaborative model[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012: 1838-1845.

[35] Yoon J H, Yang M H, Yoon K J. Interacting Multiview Tracker[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 38(5): 903-917.