2018 IEEE International Conference on Consumer Electronics (ICCE)

Simple Online and Realtime Tracking with Spherical Panoramic Camera

Keng-Chi Liu*, Yi-Ting Shen*, Liang-Gee Chen, Fellow, IEEE
DSP/IC Design Lab
National Taiwan University, Taipei, Taiwan
Email: {r05943002, r05943001, lgchen}@ntu.edu.tw

***Abstract*—In this paper, a simple yet effective method for online and realtime multi-object tracking (MOT) in 360-degree equirectangular panoramic videos is proposed. Based on the current state-of-the-art tracking-by-detection paradigm, several improvements have been made to overcome the challenges of full field-of-view (FOV) of Spherical Panoramic Camera (SPC). In addition, two datasets are presented for evaluation. It is shown that the proposed method outperforms the baseline by 28.6% and 27.8% in terms of average Multiple Object Tracking Accuracy (MOTA) on each dataset.**

***Index Terms*—spherical panoramic camera, equirectangular panorama, multi-object tracking, online and realtime tracking, convolution neural network**

I. Introduction

To create content for virtual reality (VR) applications, Spherical Panoramic Cameras (SPC) as shown in Fig. 1 are gaining in popularity due to their inexpensiveness and compact form [1]. With a full view-angle, users can interactively choose their desired viewpoint from the captured scene and have immersive, “being-there”, experience [2]. Recently, much effort has been made to further improve the processing and communicating speed for real-time 360-degree video streaming [3].

In addition to entertainment purposes, many other applications may benefit from the full field-of-view (FOV) offered by SPC, including robotic vision, autonomous vehicles, surveillance, etc. For instance, once a robot is capable of detecting and tracking all surrounding people, it can interact with them simultaneously. Due to the potential advantages above, this paper aims to extend the well-studied problem, Multi-Object Tracking (MOT) [4], especially multi-person tracking, to video sequences captured by SPC.

360-degree equirectangular panoramic videos are usually stitched and warped from two sequences captured by dual fisheye cameras as shown in Fig. 2. Although convolutional neural network (CNN) possesses convincing generalization property on scene labeling for highly distorted 360-degree video [5], there are three major difficulties when directly applying the state-of-the-art detectors [6] and trackers [7] to it. To begin with, an object that moves out of a 360-degree equirectangular panoramic image boundary will reappear at the other side immediately. Traditional trackers will lose track of this object and cause a mismatch error,or equivalently an identity switch (IDSW). Futhermore, each object size tends to be relatively small in 360-degree equirectangular panoramic images due to their full FOV. It has been shown that current detectors often led to low-precision results for tiny objects [8]. Finally, since labeling on such full FOV and highly distorted sequences is much more challenging and time-consuming, there is a lack of 360-degree MOT datasets for training and evaluation.

Fig. 1. Spherical panoramic cameras (Samsung Gear 360, Ricoh Theta S).

To overcome the problems mentioned above, a simple yet effective method for online and realtime MOT on 360-degree equirectangular panoramic videos is proposed in this paper. The overall framework is shown as Fig. 3. In order to reach better detection results, we crop out the high distorted regions of the input image and detect the remaining part in a ”multi-section” fashion. Additionally, a novel boundary handling strategy is added into the tracking pipeline to avoid IDSW. To validate the proposed method, we construct two MOT datasets, namely Basic360 and App360, and show that the proposed method outperforms the baseline by 28.6% and 27.8% in terms of average Multiple Object Tracking Accuracy (MOTA) on each dataset. We hope that this work will pave the way for new computer vision applications with SPC.

The rest of the paper is organized as follows. In section II, related work is listed and briefly introduced. The proposed approach is presented in section III. Section IV contains the analysis of the proposed datasets and experimental results. Section V highlights the applications of this work and gives some further analysis. Section VI gives our conclusions about the results of our research.

II. Related work

A. Multi-Object Tracking

Since several evaluation benchmarks were published [4] [9], many methods have been proposed for MOT. Due to recent progress in object detection [10], tracking-by-detection

[*] indicates equal contribution

978-1-5386-3025-9/18/$31.00 ©2018 IEEE

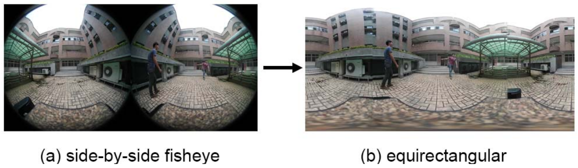


Fig. 2. (a) Two images captured by frontal and rear fisheye cameras of SPC. (b) A 360-degree equirectangular panoramic image stitched and warped from two fisheye images shown in (a).

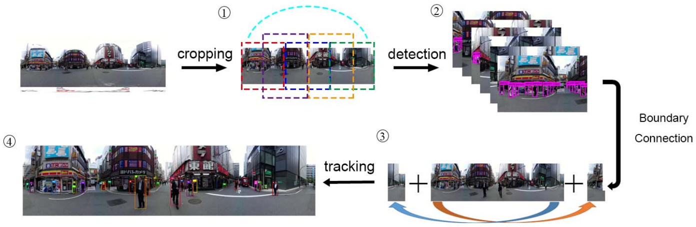


Fig. 3. Overall framework for the proposed method which includes (1) image cropping and slicing, (2) detection, (3) boundary continuity and (4) tracking.

has become the state-of-art paradigm for MOT. This kind of method can be further classified into two main categories. One usually maps MOT as a global optimization problem and process of the entire video once [11], while the other often performs data association frame by frame, which is more suitable for online scenarios [12]. In this paper, we focus on the latter.

This work is built on DeepSORT [7], an extended version of SORT [12]. Simple online and realtime tracking (SORT) achieves good performance and high speed by applying Kalman filtering and Hungarian algorithm with bounding box association measures frame by frame. DeepSORT further improves the performance by generating features of detectors through CNN which reduces the identity switch. However, none of these methods has been proven on 360-degree equirectangular panoramic sequences.

B. Spherical Panoramic Camera

Due to the rising popularity of SPC, image and video processing on 360-degree video sequences has become a hot topic. For example, Hu and Lin et al. [13] train an agent to automatically control the viewing angle on 360-degree sports videos with a recurrent neural network (RNN). Im et al. [1] extends Structure-from-Motion (SfM) to output sequences of SPC for all-around 3D reconstruction. Xu et al. [14] estimates the geometry of a room and 3D pose of objects from a single 360-degree equirectangular panoramic image.

Perhaps the most relevant work to ours is [15], which aims to track unknown objects on 360-degree polar sequences. After manually selected the desired objects in the first frame, [15] use an online training detector to detect and track them continuously in the following frames. In this paper, we aim to tackle a more challenging scenario which the proposed method should detect and track all objects of desired classes (e.g. humans) automatically. Additionally, instead of 360-degree polar sequences, the proposed method performs detection and tracking on 360-degree equirectangular panoramic sequences, since they contain relatively less distortion and the knowledge that CNN learned from normal datasets can be probably more well-transferred.

III. Approach

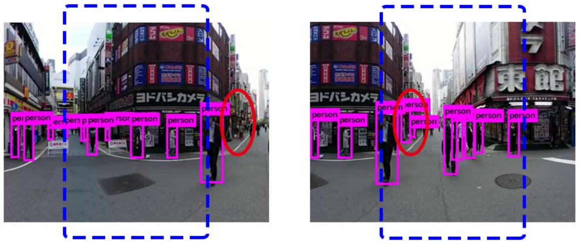
The proposed system consists of four modules as shown in Fig. 3. The first module crops and equally slices the input image into six overlapped sections. The second module then extracts 2D bounding boxes of the object class of interest (e.g. human) with a state-of-the-art detector at each section. Importantly, dectection for these six sections can be performed in parallel to maintain the original high throughput. The third module extends the cropped panoramic image for boundary continuity to prevent IDSW. Finally, an online and realtime tracker that achieves great performance at high frame rates on MOT16 benchmark [4] is adopted. The details are described below:

A. Image Cropping and Slicing

Object detection with CNN on panoramic images suffers from accuracy drops when applied to by full FOV. Lack of training image data with relatively large aspect ratio leads to bad performance. Even if the networks were specifically trained on panoramic image datasets, an inborn characteristic of panoramic images makes most objects, which are in common size in normal photos, look like small objects. Small objects detection still remains an unsolved problem nowadays [8]. Hence, to overcome the full FOV issue and to effectively fit panoramic images to the trained detector, the method described below becomes more essential. First, the input equirectangular image (1280x720) is cropped (1280x320) so as to remove the top and bottom regions which usually contain larger visual distortion and fewer interested objects (e.g. pedestrian). Then the rest part of the image is resized (2880x720) and sliced into six equally-divided sections (960x720) to get a relatively normal aspect ratio for each. It is worth mentioning that the six sections are overlapped one after another to prevent objects from appearing at the boundary between two sections as shown in step 2 of Fig. 3.

B. Detection

Considering of both accuracy and realtime requirements, the proposed method applied YOLOv2 [6] proposed by Redmon et al. as object detector. With several modifications, including dimension clustering, fully convolution neural network with anchor boxes, direct location prediction, YOLOv2 outperforms its old version YOLOv1 [16] by a large margin in terms of accuracy. Furthermore, constructed with only small size filters below 3x3, YOLOv2 becomes the fastest objector detector in current literature. Despite showing remarkable improvement, the black-box nature of neural network remains unchanged. According to our experiments, the same object appearing in the overlapped region of two neighboring sections may come out with different results when feeding each of them to YOLOv2



(a) slice (x)

(b) slice (x+1)

Fig. 4. The detection results of the same scene may differ from section to section (red). In our method, we only extract the bounding boxes which locate at the middle part of every sections (blue).

detector as shown in Fig. 4. More interestingly, the closer the object to the middle of the section, the higher accuracy we achieve. This may be due to the fact that objects in the training datasets, especially large objects, usually locates around the center of the images. Therefore, we only extract the bounding boxes which locate at the middle part of every sections. Besides, for comparison, we have also evaluated the proposed method with the most powerful object detection model, faster RCNN with inception resnet v2 [10] [17] [18] recently published by Google Inc. in section V.

C. Boundary Continuity

In addition to FOV, another major difference between panoramic images and normal photos is the continuity of leftmost and rightmost image boundary. Therefore, tracking IDs should maintain unchanged whenever tracked objects move across the image boundary. None of the existing tracking algorithms addresses this issue. To overcome this problem, we extend the cropped panoramic image for the detector as shown in Fig. 3. In this way, the trackers will not miss the corresponding detector near the image boundary. Once the tracked object moves beyond one side of the image boundary, it is considered to appear at the opposite side and the relating 2D bounding box will be shifted accordingly.

D. Tracking

As descibed in section II, the proposed method adopted DeepSORT [7], an online tracking algorithm that achieves favorable performance at high frame rates on MOT16 benchmark. It uses CNN as a deep descriptor for each bounding box that extracted by YOLOv2 and applies Kalman filtering as its tracking framework. Different from DeepSORT, we recursively generate detectors on whole extended image for matching of trackers, remove the untracked duplicated detectors after the tracking procedure, and shift the updated trackers that are out of the image boundary to their correct positions as shown in Fig. 5.

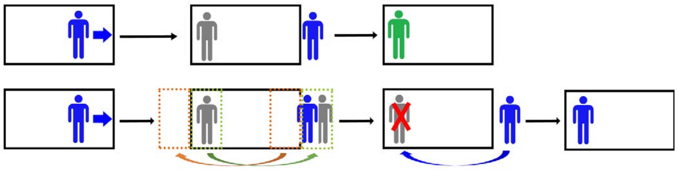


Fig. 5. Top row illustrates baseline method which will lead to IDSW (blue to green). Our method is shown at the bottom row. The flow includes generating detectors (grey) on whole extended image, removing the untracked duplicated detectors after tracking, and adjusting the trackers (remain blue).

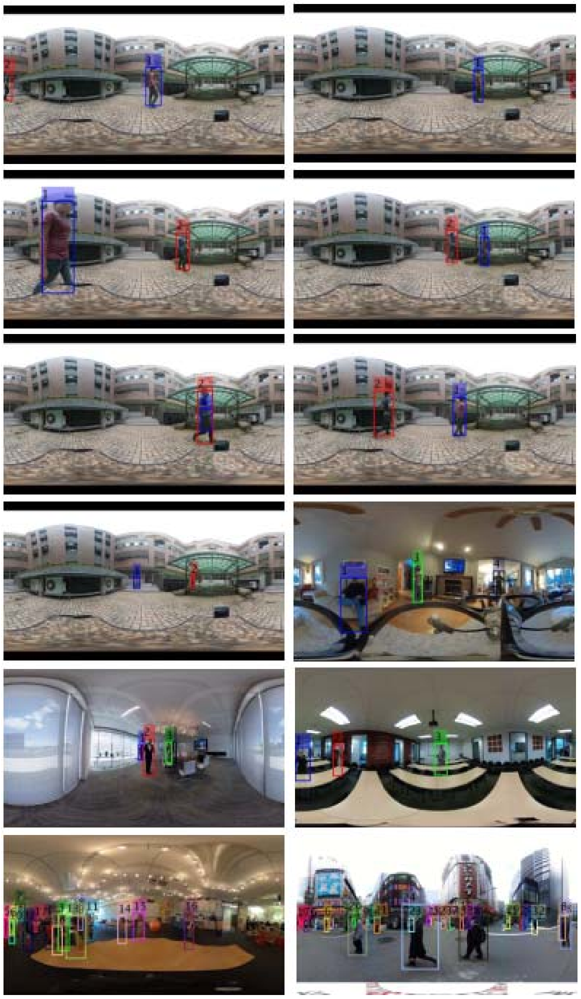


Fig. 6. The presented datasets. Left row, from top to bottom: boundary, distance, overlap, sit, microsoft, kindergarten. Right row, from top to bottom: back, jump, run, drone, meeting, japan.

IV. Experiment

A. Dataset Description

To evaluate the proposed method, twelve 360-degree equirectangular panoramic sequences have been collected (Fig. 6) and carefully annotated by following a consistent protocol suggested in [4]. The annotation tool ViTBAT proposed by [19] has been applied to facilitate the process. The collected sequences have been divided into two evaluation datasets, namely Basic360 and App360. Details are listed in Table. I and Table. II. Briefly, Basic360 is a simple dataset that aims to prove the correctness of the proposed method. We create this dataset which contains only two men performing several basic actions with Samsung Gear 360. On the other hand, App360 targets some potential applications for MOT with SPC (see section V for more details) and contains more challenging and realistic scenes.

TABLE I
THE PRESENTED BASIC360 DATASET.

Basic360 dataset					
Name	FPS	Resolution	Length	Annotated Humans	Description
boundary	30	1280x720	282	2	Men walk across the image boundary
back	30	1280x720	222	2	A man turns his back to face camera
distance	30	1280x720	270	2	A man stands near the camera
jump	30	1280x720	90	2	A man jumps in front of the camera
overlap	30	1280x720	96	2	A man stands behind the other man
run	30	1280x720	132	2	Men run around the camera
sit	30	1280x720	408	2	A man sits on the stairs

TABLE II
THE PRESENTED APP360 DATASET.

App360 dataset					
Name	FPS	Resolution	Length	Annotated Humans	Description
drone	30	1280x720	1234	4	A drone with a SPC flies around in the house
microsoft	30	1280x720	126	4	People walks into the meeting room
meeting	30	1280x720	449	4	People walks around in the meeting room
kindergarten	30	1280x720	378	18	Many children walks around in the classroom
japan	30	1280x720	616	45	Pedestrians in Japan

B. Baseline Method and Evaluation Metrics

To highlight the effect of the proposed method, we set the original DeepSORT [7] as baseline. For fair comparison, both DeepSORT and the proposed method applied the same object detector, YOLOv2 [6]. Additionally, to reduce the negative impact caused by image distortion, both of them perform detection and tracking on the 360-degree equirectangular panoramic images after cropping out the top and bottom regions as described in section III. The only two differences between the two methods are that the proposed method performs detection on six overlapped image sections and has a strategy of handling boundary continuity issue as described in section III.

For evaluation, we adopt the public tools and evaluation metrics provided by the MOT16 Benchmark. More details can be found in their website[1].

C. Evaluation Results on Basic360 dataset

The evaluation results for Basic360 dataset are shown in Table. III and Table. IV. The proposed method outperforms the baseline in many aspects. For example, it is shown that the proposed boundary handling method effectively reduces the total identity switch (IDSW). Moreover, it is also shown that the proposed image slicing method decreases the total number of false positives (FP) and false negatives (FN) by a large amount. As a result, the proposed method outperforms the baseline by 28.6% in terms of Multi-Object Tracking Accuracy (MOTA).

Despite the informal improvement achieved by the proposed method, it performs relatively poor on "sit" and "back" sequences. Both of these sequences contain scenes which the targets are too far from the camera. As mentioned in section III, current CNN-based detectors often have worse performance on tiny objects. Such poor detection quality will result in bad tracking performance for these scenes.

V. DISCUSSION

A. Evaluation Results on App360 dataset

To addressed the potential applications for 360-degree MOT with SPC, we test the proposed method on App360 dataset. This dataset mainly targets the following four applications: Pedestrian Monitoring ("japan" sequence), Automatic Babysitting ("kindergarten" sequence), Interactive Meeting ("meeting" sequence, "microsoft" sequence) and Autonomous Drone ("drone" sequence).

The evaluation results on App360 dataset are shown in Table. V and Table. VI. Although the performance is much worse than the performance on Basic360 due to more complicated scenarios, the proposed method still largely outperforms baseline by 27.8% in terms of MOTA.

B. Sensitivity to Detection Accuracy

According to Table. VI, the poor performance of the proposed method basically arises from large number of false positives and false negatives. Despite the goal of this research online and realtime tracking, we were curious as to whether the tracking performance could be further improved by using a more accurate but slower detector. To answer this question, we replace YOLOv2 in the proposed method with the faster RCNN with inception resnet v2 implemented by [10], which recently achieved the best detection accuracy in the COCO challenge [20]. The results are shown in Table. VII. The proposed method with a state-of-the-art detector outperforms the original method by 21.4% in terms of MOTA. Once again, the result implies that improvement in detection quality can further enhance the tracking performance by a large margin.

VI. CONCLUSION

Recently, demand for high-performance object detection and tracking has grown drastically due to their potentially high commercial opportunities. In this paper, the first online and realtime multi-object tracking algorithm on 360-degree

[1] https://motchallenge.net/

TABLE III

The Baseline method on Basic360 dataset.

The Baseline Method																
Name	IDF1	IDP	IDR	Rcll	Prcn	FAR	GT	MT	PT	ML	FP	FN	IDSW	FM	MOTA	MOTP	MOTAL
boundary	85.7	84.4	87.1	97.3	94.3	0.12	2	2	0	0	33	15	2	1	91.1	74.4	91.4
back	56.8	58.5	55.2	66.4	70.4	0.56	2	1	1	0	124	149	5	6	37.4	67.6	38.8
distance	89.9	91.1	88.7	89.8	92.2	0.15	2	2	0	0	41	55	1	1	82.0	69.7	82.2
jump	81.3	79.4	83.3	84.4	80.4	0.41	2	1	1	0	37	28	1	1	63.3	67.0	63.7
overlap	94.3	98.3	90.6	90.6	98.3	0.03	2	2	0	0	3	18	0	2	89.1	69.9	89.1
run	58.9	57.3	60.6	84.5	79.9	0.42	2	1	1	0	56	41	5	5	61.4	73.1	63.0
sit	20.5	25.4	17.2	38.2	56.5	0.59	2	0	1	1	240	504	5	17	8.2	56.0	8.7
Total	33.4	35.1	31.8	73.0	80.4	0.36	14	9	4	1	534	810	19	33	54.6	68.8	55.2

TABLE IV

The Proposed method on Basic360 dataset

The Proposed Method																
Name	IDF1	IDP	IDR	Rcll	Prcn	FAR	GT	MT	PT	ML	FP	FN	IDSW	FM	MOTA	MOTP	MOTAL
boundary	99.2	99.6	98.8	98.8	99.6	0.01	2	2	0	0	2	7	0	2	98.4	78.6	98.4
back	82.9	83.4	82.4	82.4	83.4	0.33	2	1	1	0	73	78	0	2	66.0	78.8	66.0
distance	99.3	99.8	98.9	98.9	99.8	0.00	2	2	0	0	1	6	0	2	98.7	74.9	98.7
jump	98.9	100.0	97.8	97.8	100.0	0.00	2	2	0	0	0	4	0	0	97.8	76.3	97.8
overlap	97.6	100.0	95.3	95.3	100.0	0.00	2	2	0	0	0	9	0	1	95.3	77.7	95.3
run	91.0	89.7	92.4	92.4	89.7	0.21	2	2	0	0	28	20	0	3	81.8	78.4	81.8
sit	69.5	75.8	64.2	75.7	89.4	0.18	2	1	0	0	73	198	4	17	66.3	76.5	66.7
Total	46.1	47.3	45.0	89.3	93.8	0.12	14	12	1	0	177	322	4	27	83.2	77.2	83.3

TABLE V

The Baseline method on App360 dataset

The Baseline Method																
Name	IDF1	IDP	IDR	Rcll	Prcn	FAR	GT	MT	PT	ML	FP	FN	IDSW	FM	MOTA	MOTP	MOTAL
drone	16.6	17.5	15.8	42.4	46.8	1.47	4	0	4	0	1816	2175	36	78	-6.7	63.7	-5.8
microsoft	41.9	59.0	32.5	44.2	80.2	0.44	4	1	2	1	55	281	2	8	32.9	65.4	33.2
meeting	15.9	28.3	11.0	14.9	38.1	0.96	4	0	1	3	433	1529	5	13	-9.5	60.7	-9.3
kindergarten	12.2	20.5	8.7	12.8	30.1	5.28	18	0	4	14	1997	5865	34	55	-17.4	58.0	-17.0
japan	9.9	31.2	5.9	8.7	46.1	3.36	45	0	4	41	2067	18561	42	95	-1.7	65.3	-1.5
Total	9.8	19.5	6.5	14.2	42.5	2.27	75	1	15	59	6368	28411	119	249	-5.3	63.2	-5.0

TABLE VI

The proposed Method on App360 dataset

The Proposed Method																
Name	IDF1	IDP	IDR	Rcll	Prcn	FAR	GT	MT	PT	ML	FP	FN	IDSW	FM	MOTA	MOTP	MOTAL
drone	30.4	28.7	32.3	60.8	53.9	1.59	4	1	3	0	1963	1479	21	88	8.2	70.1	8.8
microsoft	68.3	68.0	68.7	84.5	83.7	0.66	4	3	1	0	83	78	2	3	67.7	74.1	68.0
meeting	48.9	40.9	60.7	76.2	51.4	2.88	4	2	2	0	1294	428	3	26	4.0	4.0	4.1
kindergarten	45.4	43.0	48.0	61.8	55.4	8.85	18	4	13	1	3344	2570	61	163	11.1	11.1	12.0
japan	37.1	45.4	31.3	49.6	71.7	6.44	45	7	29	9	3967	10254	132	309	29.4	29.4	30.0
Total	27.8	29.8	26.1	55.3	63.2	3.80	75	17	48	10	10651	14809	219	589	22.5	22.5	23.1

TABLE VII

State-of-the-art detector with the proposed method.

The Proposed method with state-of-art detector																
Name	IDF1	IDP	IDR	Rcll	Prcn	FAR	GT	MT	PT	ML	FP	FN	IDSW	FM	MOTA	MOTP	MOTAL
drone	42.2	40.0	44.6	68.2	61.1	1.33	4	1	3	0	1640	1199	20	66	24.2	69.8	24.7
microsoft	71.3	73.1	69.6	82.5	86.7	0.51	4	2	2	0	64	88	2	5	69.4	73.9	69.7
meeting	91.8	93.2	90.5	90.5	93.2	0.27	4	4	0	0	119	170	0	14	83.9	69.9	83.9
kindergarten	50.9	53.0	49.0	64.0	69.3	5.04	18	5	12	1	1907	2417	48	127	35.0	66.8	35.7
japan	47.3	52.7	42.9	64.2	79.0	5.64	45	10	32	3	3474	7273	152	352	46.4	71.3	47.1
Total	34.5	36.8	32.4	66.3	75.3	2.57	75	22	49	4	7204	11147	222	564	43.9	70.2	44.6

equirectangular panoramic sequences is proposed. It is shown that the proposed method outperforms the baseline by a large margin. We believe that solving computer vision problems with SPCs can not only makes the original system (e.g. robotics, autonomous vehicle and surveillance) more powerful but also provides a more cost-effective solution when compared with multi-cameras system. For future works, we plan to solve other computer vision problems with spherical panoramic cameras, such as depth and surface normal estimation.

References

[1] S. Im, H. Ha, F. Rameau, H.-G. Jeon, G. Choe, and I. S. Kweon, "All-around depth from small motion with a spherical panoramic camera," in *European Conference on Computer Vision*. Springer, 2016, pp. 156–172.

[2] T.-M. Liu, C.-C. Ju, Y.-H. Huang, T.-S. Chang, K.-M. Yang, and Y.-T. Lin, "A 360-degree 4k-2k panoramic video processing over smart-phones," in *2017 IEEE International Conference on Consumer Electronics (ICCE)*, Jan 2017, pp. 247–249.

[3] "Facebook live 360." [Online]. Available: https://facebook360.fb.com/live360/

[4] A. Milan, L. Leal-Taixé, I. D. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," *CoRR*, vol. abs/1603.00831, 2016. [Online]. Available: http://arxiv.org/abs/1603.00831

[5] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1915–1929, 2013.

[6] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," *CoRR*, vol. abs/1612.08242, 2016. [Online]. Available: http://arxiv.org/abs/1612.08242

[7] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," *CoRR*, vol. abs/1703.07402, 2017. [Online]. Available: http://arxiv.org/abs/1703.07402

[8] C. Chen, M.-Y. Liu, O. Tuzel, and J. Xiao, *R-CNN for Small Object Detection*. Cham: Springer International Publishing, 2017, pp. 214–230.

[9] L. Leal-Taixé, A. Milan, I. D. Reid, S. Roth, and K. Schindler, "Motchallenge 2015: Towards a benchmark for multi-target tracking," *CoRR*, vol. abs/1504.01942, 2015. [Online]. Available: http://arxiv.org/abs/1504.01942

[10] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy, "Speed/accuracy trade-offs for modern convolutional object detectors," *CoRR*, vol. abs/1611.10012, 2016. [Online]. Available: http://arxiv.org/abs/1611.10012

[11] S. Tang, B. Andres, M. Andriluka, and B. Schiele, "Multi-person tracking by multicut and deep matching," *CoRR*, vol. abs/1608.05404, 2016. [Online]. Available: http://arxiv.org/abs/1608.05404

[12] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," *CoRR*, vol. abs/1602.00763, 2016. [Online]. Available: http://arxiv.org/abs/1602.00763

[13] H. Hu, Y. Lin, M. Liu, H. Cheng, Y. Chang, and M. Sun, "Deep 360 pilot: Learning a deep agent for piloting through 360-degree sports video," *CoRR*, vol. abs/1705.01759, 2017. [Online]. Available: http://arxiv.org/abs/1705.01759

[14] J. Xu, B. Stenger, T. Kerola, and T. Tung, "Pano2cad: Room layout from A single panorama image," *CoRR*, vol. abs/1609.09270, 2016. [Online]. Available: http://arxiv.org/abs/1609.09270

[15] A. Delforouzi, S. A. H. Tabatabaei, K. Shirahama, and M. Grzegorzek, "Unknown object tracking in 360-degree camera images," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, Dec 2016, pp. 1798–1803.

[16] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *CoRR*, vol. abs/1506.02640, 2015. [Online]. Available: http://arxiv.org/abs/1506.02640

[17] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015. [Online]. Available: http://arxiv.org/abs/1506.01497

[18] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, inception-resnet and the impact of residual connections on learning," *CoRR*, vol. abs/1602.07261, 2016. [Online]. Available: http://arxiv.org/abs/1602.07261

[19] T. A. Biresaw, T. Nawaz, J. Ferryman, and A. I. Dell, "Vitbat: Video tracking and behavior annotation tool," in *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Aug 2016, pp. 295–301.

[20] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.